2022

FRM®

EXAM PART I

Quantitative Analysis

GARP

FRM | Financial Risk Manager

VISIT…

LANZAROTE Caliente.com

GARP
FRM Financial Risk Manager

2022

FRM®

EXAM PART I

Quantitative Analysis

Pearson

Copyright © 2022, 2021 by the Global Association of Risk Professionals
All rights reserved.

This copyright covers material written expressly for this volume by the editor/s as well as the compilation itself. It does not cover the individual selections herein that first appeared elsewhere. Permission to reprint these has been obtained by Pearson Education, Inc. for this edition only. Further reproduction by any means, electronic or mechanical, including photocopying and recording, or by any information storage or retrieval system, must be arranged with the individual copyright holders noted.

All trademarks, service marks, registered trademarks, and registered service marks are the property of their respective owners and are used herein for identification purposes only.

Pearson Education, Inc., 330 Hudson Street, New York, New York 10013
A Pearson Education Company
www.pearsoned.com

Printed in the United States of America

1 2 3 4 5 6 7 8 9 10 XXXX 19 18 17 16

00006613-00000002 / A103000247797

EEB/MB

Pearson

ISBN 10: 0137440979
ISBN 13: 9780137440979

Contents

Chapter 1 **Fundamentals of Probability** 1

1.1 Sample Space, Event Space, and Events 2
Probability 2
Fundamental Principles of Probability 4
Conditional Probability 4
Example: SIFI Failures 5
1.2 Independence 5
Conditional Independence 6
1.3 Bayes' Rule 7
1.4 Summary 7
Questions 8
Short Concept Questions 8
Practice Questions 8
Answers 9
Short Concept Questions 9
Solved Problems 9

Chapter 2 **Random Variables** 11

2.1 Definition of a Random Variable 12
2.2 Discrete Random Variables 13
2.3 Expectations 14
Properties of the Expectation Operator 14
2.4 Moments 14
The Four Named Moments 16
Moments and Linear Transformations 17
2.5 Continuous Random Variables 18
2.6 Quantiles and Modes 20
2.7 Summary 22
Questions 23
Short Concept Questions 23
Practice Questions 23
Answers 24
Short Concept Questions 24
Solved Problems 25

Chapter 3 Common Univariate Random Variables 27

3.1 Discrete Random Variables 28
Bernoulli 28
Binomial 29
Poisson 30
3.2 Continuous Random Variables 31
Uniform 31
Normal 32
Approximating Discrete Random Variables 34
Log-Normal 34
χ^2 36
Student's t 37
F 38
Exponential 39
Beta 40
Mixtures of Distributions 40
3.3 Summary 43
Questions 44
Short Concept Questions 44
Practice Questions 44
Answers 45
Short Concept Questions 45
Solved Problems 45

Chapter 4 Multivariate Random Variables 47

4.1 Discrete Random Variables 48
Probability Matrices 48
Marginal Distributions 49
Independence 50
Conditional Distributions 50
4.2 Expectations and Moments 51
Expectations 51
Moments 52
The Variance of Sums of Random Variables 53
Covariance, Correlation, and Independence 54
4.3 Conditional Expectations 55
Conditional Independence 55
4.4 Continuous Random Variables 56
Marginal and Conditional Distributions 57
4.5 Independent, Identically Distributed Random Variables 57
4.6 Summary 58
Questions 59
Short Concept Questions 59
Practice Questions 59
Answers 60
Short Concept Questions 60
Solved Problems 60

Chapter 5 Sample Moments 63

5.1 The First Two Moments 64
Estimating the Mean 64
Estimating the Variance and Standard Deviation 65
Presenting the Mean and Standard Deviation 66
Estimating the Mean and Variance Using Data 67
5.2 Higher Moments 67
5.3 The BLUE Mean Estimator 70
5.4 Large Sample Behavior of the Mean 71

5.5 The Median and Other Quantiles 73
5.6 Multivariate Moments 74
Covariance and Correlation 74
Sample Mean of Two Variables 75
Coskewness and Cokurtosis 76
5.7 Summary 78
Questions 79
Short Concept Questions 79
Practice Questions 79
Answers 80
Short Concept Questions 80
Solved Problems 80

Chapter 6 Hypothesis Testing 83

6.1 Elements of a Hypothesis Test 84
Null and Alternative Hypotheses 84
One-Sided Alternative Hypotheses 85
Test Statistic 85
Type I Error and Test Size 86
Critical Value and Decision Rule 86
Example: Testing a Hypothesis about the Mean 86
Type II Error and Test Power 88
Example: The Effect of Sample Size on Power 89
Example: The Effect of Test Size and Population Mean on Power 90
Confidence Intervals 90
The p-value of a Test 91
6.2 Testing the Equality of Two Means 93
6.3 Summary 94
Questions 95
Short Concept Questions 95
Practice Questions 95
Answers 96
Short Concept Questions 96
Solved Problems 96

Chapter 7 Linear Regression 101

7.1 Linear Regression 102
Linear Regression Parameters 102
Linearity 103
Transformations 103
Dummy Variables 104
7.2 Ordinary Least Squares 104
7.3 Properties of OLS Parameter Estimators 106
Shocks Are Mean Zero 106
Data Are Realizations from iid Random Variables 107
Variance of X 108
Constant Variance of Shocks 108
No Outliers 108
Implications of OLS Assumptions 109
7.4 Inference and Hypothesis Testing 110
7.5 Application: CAPM 111
7.6 Application: Hedging 113
7.7 Application: Performance Evaluation 114
7.8 Summary 116
Questions 117
Short Concept Questions 117
Practice Questions 117
Answers 118
Short Concept Questions 118
Solved Problems 118

Chapter 8 Regression with Multiple Explanatory Variables 121

8.1 Multiple Explanatory Variables 122
Additional Assumptions 123
8.2 Application: Multi-Factor Risk Models 124
8.3 Measuring Model Fit 127
8.4 Testing Parameters In Regression Models 129
Multivariate Confidence Intervals 130
The *F*-statistic of a Regression 133
8.5 Summary 133
8.6 Appendix 134
Tables 134
Questions 136
Short Concept Questions 136
Practice Questions 136
Answers 137
Short Concept Questions 137
Solved Problems 137

Chapter 9 Regression Diagnostics 141

9.1 Omitted Variables 142
9.2 Extraneous Included Variables 143
9.3 The Bias-Variance Tradeoff 143
9.4 Heteroskedasticity 144
Approaches to Modeling Heteroskedastic Data 147
9.5 Multicollinearity 147
Residual Plots 148
Outliers 148
9.6 Strengths of OLS 149
9.7 Summary 151
Questions 152
Short Concept Questions 152
Practice Questions 152
Answers 155
Short Concept Questions 155
Solved Problems 155

Chapter 10 Stationary Time Series 161

10.1 Stochastic Processes 162
10.2 Covariance Stationarity 163
10.3 White Noise 164
Dependent White Noise 165
Wold's Theorem 165
10.4 Autoregressive (AR) Models 165
Autocovariances and Autocorrelations 166
The Lag Operator 166
AR(p) 167
10.5 Moving Average (MA) Models 171
10.6 Autoregressive Moving Average (ARMA) Models 173
10.7 Sample Autocorrelation 176
Joint Tests of Autocorrelations 176
Specification Analysis 176
10.8 Model Selection 179
Box-Jenkins 180

10.9 Forecasting 180
10.10 Seasonality 183
10.11 Summary 184
Appendix 184
Characteristic Equations 184
Questions 185
Short Concept Questions 185
Practice Questions 185
Answers 186
Short Concept Questions 186
Solved Problems 186

Chapter 11 Non-Stationary Time Series 189

11.1 Time Trends 190
Polynomial Trends 190
11.2 Seasonality 192
Seasonal Dummy Variables 193
11.3 Time Trends, Seasonalities, and Cycles 193
11.4 Random Walks and Unit Roots 196
Unit Roots 196
The Problems with Unit Roots 196
Testing for Unit Roots 197
Seasonal Differencing 200
11.5 Spurious Regression 201
11.6 When to Difference? 202
11.7 Forecasting 203
Forecast Confidence Intervals 204
11.8 Summary 204
Questions 206
Short Concept Questions 206
Practice Questions 206
Answers 208
Short Concept Questions 208
Solved Problems 208

Chapter 12 Measuring Returns, Volatility, and Correlation 213

12.1 Measuring Returns 214
12.2 Measuring Volatility and Risk 215
Implied Volatility 216
12.3 The Distribution of Financial Returns 216
Power Laws 217
12.4 Correlation Versus Dependence 218
Alternative Measures of Correlation 218
12.5 Summary 221
Questions 222
Short Concept Questions 222
Practice Questions 222
Answers 223
Short Concept Questions 223
Solved Problems 223

Chapter 13 Simulation and Bootstrapping 225

13.1 Simulating Random Variables 226
13.2 Approximating Moments 226

The Mean of a Normal 228
Approximating the Price of a Call Option 229
Improving the Accuracy of Simulations 231
Antithetic Variates 231
Control Variates 231
Application: Valuing an Option 232
Limitation of Simulations 233

13.3 Bootstrapping 233

Bootstrapping Stock Market Returns 234
Limitations of Bootstrapping 236

13.4 Comparing Simulation and Bootstrapping 236

13.5 Summary 236

Questions 237

Short Concept Questions 237
Practice Questions 237

Answers 238

Short Concept Questions 238
Solved Problems 238

Index 241

PREFACE

On behalf of our Board of Trustees, GARP's staff, and particularly its certification and educational programs teams, I would like to thank you for your interest in and support of our Financial Risk Manager (FRM®) program.

The past couple of years have been difficult due to COVID-19. And in that regard, our sincere sympathies go out to anyone who was ill or suffered a loss due to the pandemic.

The FRM program also experienced many virus-related challenges. Because we always place candidate safety first, we cancelled the May 2020 FRM exam offering and deferred all candidates to October, while reserving an optional date in January 2021 for candidates not able to sit for the examination in October. A change like this has never happened before. Ultimately, we were able to offer the FRM exam to all 2020 registered candidates who wanted to sit for it during the year and were not constrained by COVID-related restrictions, which was most of our registrants.

Since its inception in 1997, the FRM program has been the global industry benchmark for risk-management professionals wanting to demonstrate objectively their knowledge of financial risk-management concepts and approaches. Having FRM holders on staff also tells companies' that their risk-management professionals have achieved a demonstrated and globally adopted level of expertise.

In a world where risks are becoming more complex daily due to any number of technological, capital, governmental, geopolitical, or other factors, companies know that FRM holders possess the skills to understand and adapt to a dynamic and rapidly changing financial environment.

The FRM program addresses the financial risks faced by both non-financial firms and those in the highly interconnected and sophisticated financial services industry, because its coverage is not static, but vibrant and forward looking.

The FRM curriculum is regularly reviewed by an oversight committee of highly qualified and experienced risk-management professionals from around the globe. These professionals consist of senior bank and consulting practitioners, government regulators, asset managers, insurance risk professionals, and academics. Their mission is to ensure the FRM program remains current and its content addresses not only standard credit and market risk issues, but also emerging issues and trends, ensuring FRM candidates are aware of what is or is expected to be important in the near future. We're committed to offering a program that reflects the dynamic and sophisticated nature of the risk-management profession and those who are making it a career.

We wish you the very best as you study for the FRM exams, and in your career as a risk-management professional.

Yours truly,

Richard Apostolik

President & CEO

FRM® COMMITTEE

Chairperson

Michelle McCarthy Beck
Former GARP Board Member

Members

Richard Apostolik
President and CEO, Global Association of Risk Professionals

Richard Brandt
MD Operational Risk Management, Citigroup

Julian Chen, FRM, SVP
FRM Program Manager, Global Association of Risk Professionals

Dr. Christopher Donohue, MD
GARP Benchmarking Initiative, Global Association of Risk Professionals

Donald Edgar, FRM
MD Risk & Quantitative Analysis, BlackRock

Hervé Geny
Former Group Head of Internal Audit, London Stock Exchange Group and former CRO, ICAP

Keith Isaac, FRM
VP Capital Markets Risk Management, TD Bank Group

William May, SVP
Global Head of Certifications and Educational Programs, Global Association of Risk Professionals

Dr. Attilio Meucci, CFA
Founder, ARPM

Dr. Victor Ng
MD Head of Risk Architecture, Goldman Sachs

Dr. Matthew Pritsker
Senior Financial Economist and Policy Advisor / Supervision, Regulation, and Credit, Federal Reserve Bank of Boston

Dr. Samantha Roberts, FRM
SVP Balance Sheet Analytics & Modeling, PNC Bank

Dr. Til Schuermann
Partner, Oliver Wyman

Nick Strange
Senior Technical Advisor, Operational Risk & Resilience, Supervisory Risk Specialists, Prudential Regulation Authority, Bank of England

Dr. Sverrir Þorvaldsson, FRM
Senior Quant, SEB

ATTRIBUTIONS

Author

Kevin Sheppard, PhD, Associate Professor in Financial Economics and Tutorial Fellow in Economics at the University of Oxford

Contributor

Chris Brooks, PhD, Professor of Finance at the ICMA Centre, University of Reading

Reviewers

Paul Feehan, PhD, Professor of Mathematics and Director of the Masters of Mathematical Finance Program, Rutgers University

Erick W. Rengifo, PhD, Associate Professor of Economics, Fordham University

Richard Morrin, PhD, FRM, ERP, CFA, CIPM, Financial Officer, International Finance Corporation

Antonio Firmo, FRM, Senior Audit Manager—Trading Risk, Risk Capital, and Modeling, Royal Bank of Scotland

Deco Caigu Liu, FRM, CSC, CBAP, Global Asset Liability Management, Manulife

John Craig Anderson, FRM, Director—Debt and Capital Advisory Services, Advize Capital

Thierry Schnyder, FRM, CSM, Risk Analyst, Bank for International Settlements

Yaroslav Nevmerzhytskyi, FRM, CFA, ERP, Deputy Chief Risk Officer and Financial Risk Management Team Leader, Naftogaz of Ukraine

1

Fundamentals of Probability

Learning Objectives

After completing this reading you should be able to:

- Describe an event and an event space.
- Describe independent events and mutually exclusive events.
- Explain the difference between independent events and conditionally independent events.
- Calculate the probability of an event for a discrete probability function.
- Define and calculate a conditional probability.
- Distinguish between conditional and unconditional probabilities.
- Explain and apply Bayes' rule.

Probability theory is the foundation of statistics, econometrics, and risk management. This chapter introduces probability in its simplest form. A probability measures the likelihood that some event occurs. In most financial applications of probability, events are tightly coupled with numeric values. Examples of such events include the loss on a portfolio, the number of defaults in a mortgage pool, or the sensitivity of a portfolio's value to a rise in short-term interest rates. Events can also be measured by values without a natural numeric correspondence. These include categorical variables, such as the type of a financial institution or the rating on a corporate bond.

Probability is introduced through three fundamental principles.

1. The probability of any event is non-negative.
2. The sum of the probabilities across all outcomes is one.
3. The joint probability of two independent events is the product of the probability of each.

This chapter also introduces conditional probability, an important concept that assigns probabilities within a subset of all events. After that, the chapter moves on to discuss independence and conditional independence. It concludes by examining Bayes' rule, which provides a simple yet powerful expression for incorporating new information or data into probability estimates.

1.1 SAMPLE SPACE, EVENT SPACE, AND EVENTS

A sample space is a set containing all possible outcomes of an experiment and is denoted as Ω. The set of outcomes depends on the problem being studied. For example, when examining the returns on the S&P 500, the sample space is the set of all real numbers and is denoted by $\mathbb{R}$.[1] On the other hand, if the focus is on the *direction* of the return on the S&P 500, then the sample space is {Positive, Negative}. When modeling corporate defaults, the sample space is {Default, No Default}. When rolling a single six-sided die, the sample space is {1, 2, . . . , 6}. The sample space for two identical six-sided dice contains 21 elements representing all distinct pairs of values: {(1,1), (1,2), (1,3), . . . , (5,5), (5,6), (6,6)}.[2] If we are modeling the sum of two dice, however, then the sample space is {2, . . . , 12}.

[1] The sample space in this example is truncated at the lower end since the return cannot be less than −100%.

[2] This assumes that the dice are indistinguishable and therefore one can only observe the values shown (and not which die produced each value). If the dice were distinguishable from each other (e.g., they have different colors), then the sample space would contain 36 values (because {1,2} and {2,1} are different).

Events are subsets of a sample space, and an event is denoted by ω. An event is a set of outcomes and may contain one or more of the values in the sample space, or it may even contain no elements.[3] When an event includes only one outcome, it is often referred to as an elementary event. Events are sets and are usually written with set notation. For example, when interested in the sum of two dice, one event is that the sum is odd: {(1,2), (1,4), (1,6), . . . , (4,5), (5,6)}. Another event is that the sum is greater than or equal to nine: {(4,5), (4,6), (5,5), (5,6), (6,6)}.

The event space, usually denoted by $\mathcal{F}$, consists of all combinations of outcomes to which probabilities can be assigned. Note that the event space is an abstract concept and separate from any specific application. As an example, suppose that {*A*}, {*B*} are the possible outcomes of an experiment. Without knowing what {*A*}, {*B*} mean in practice, one might consider the following events.

1. *A* occurs, *B* does not.
2. *B* occurs, *A* does not.
3. Both *A* and *B* occur.
4. Neither *A* nor *B* occur.

This would give an event space consisting of four events {*A*, *B*, {*A*, *B*}, ∅}. Because this event space has a finite number of outcomes, it is called a *discrete probability space*.[4]

In the corporate default example, the event space is {Default, No Default, {Default, No Default}, ∅}. It might appear surprising that the event space contains two "impossible" events for this example:

1. {Default, No Default}, which contains both outcomes; and
2. The empty set ∅, which contains neither.

However, note that a definite probability of 0 can be assigned to "impossible" events. Because the event space contains all sets that can be assigned a probability, these two events are therefore part of the event space.

Probability

Probability measures the likelihood of an event. Probabilities are always between 0 and 1 (inclusive). An event with probability 0 never occurs, while an event with a probability 1 always occurs. The simplest interpretation of probability is that it is the frequency with which an event would occur if a set of independent experiments was run. This interpretation of probability is known as the frequentist interpretation and focuses on objective

[3] An event that contains no elements is known as the *empty set*.

[4] A discrete probability space is one that contains either finitely many outcomes or (possibly) countably infinitely many outcomes.

probability. Note that this is a conceptual interpretation, whereas finance is mostly focused on non-experimental events (e.g., the return on the stock market).

Probability can also be interpreted from a subjective point of view. Under this alternative interpretation, probability reflects or incorporates an individual's beliefs about the likelihood of an event occurring. These beliefs may differ across individuals and do not have to agree with the objective probability of an event. In fact, there might not even be a single objective probability that can be assigned to each outcome in many real-world situations if the true model or process generating the outcomes is unknown. An example of this would be a senior executive assuming a specific probability for an event (e.g., a recession, an interest rate increase, or a rating downgrade) as part of a scenario analysis or stress test.

Probability is defined over event spaces. Mathematically, it assigns a number between 0 and 1 (inclusive) to each event in the event space. As stated above, events are simply combinations of outcomes (i.e., subsets of the sample space). It is therefore possible to illustrate some core probability concepts using the language of sets.

Figure 1.1 illustrates the key properties of sets, with each rectangle representing a sample space (Ω) comprised of two events (i.e., A and B). Note that the standard notation for a set uses capital letters (e.g., A, B), while subscripts (e.g., A_i) are used to denote subsets. Three important set operators shown in the figure are intersections, unions, and complements:

- $\cap$ is the intersection operator, which means that $A \cap B$ is the set of outcomes that appear in both A and B.
- $\cup$ is the union operator, which means that $A \cup B$ is the set that contains all outcomes in A, or B, or both.
- A^c is the complement of the set A, and is the set of all outcomes that are not in A.

Finally, the bottom-right panel illustrates mutually exclusive events, which means that either A or B could happen, but not both together. Here, the intersection $A \cap B$ is the empty set ($\varnothing$).

Recall the earlier example of rolling two dice, which featured two events:

1. The event that the sum rolled is greater than or equal to nine; and
2. The event that the sum is odd.

The intersection of these two events contains the outcomes from the sample space where both conditions are simultaneously satisfied. This happens when the sum is equal to nine or 11, meaning that the intersection of these two events consists of the outcomes {(4,5), (5,6)}. The union of the two events includes all outcomes in both sets without duplication (i.e., the outcomes in the set of odd sums as well as the even sums greater than or equal to nine). The

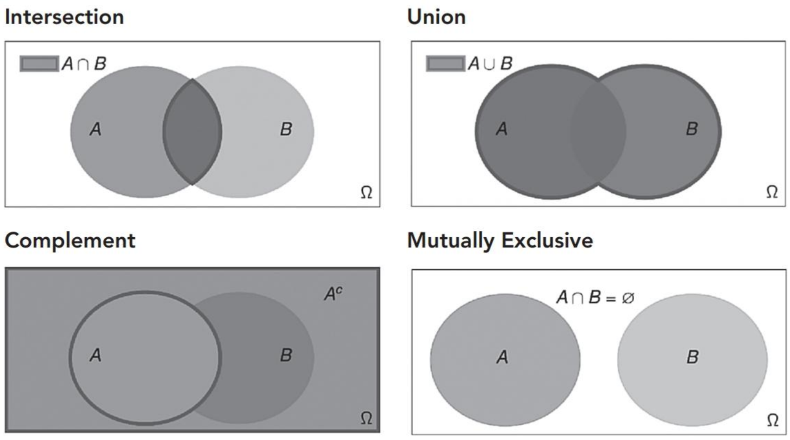


Figure 1.1 **The top-left panel illustrates the intersection of two sets, which is indicated by the shaded area surrounded by the dark solid line. The top right panel illustrates the union of two sets, which is indicated by the shaded area surrounded by the dark solid line. The shaded area in the bottom-left panel illustrates A^c, which is the complement to A and includes every outcome that is not in A. The bottom-right panel illustrates mutually exclusive sets, which have an empty intersection (denoted by $\varnothing$).**

complement of the event that the sum is odd is the set of outcomes where the sum of two dice is even: ({1,1}, {1,3}, {1,5}, . . . , (5,5), (6,6)}. Finally, the intersection of the event that the sum is odd with the event that the sum is even is the null set (because there cannot be a sum which is simultaneously odd and even). This is an example of two mutually exclusive events.

Fundamental Principles of Probability

The three fundamental principles of probability are defined using events, event spaces, and sample spaces. Note that $\Pr(\omega)$ is a function that returns the probability of an event ω.

1. Any event A in the event space $\mathcal{F}$ has $\Pr(A) \geq 0$.
2. The probability of all events in Ω is one and thus $\Pr(\Omega) = 1$.
3. If the events A_1 and A_2 are mutually exclusive, then $\Pr(A_1 \cup A_2) = \Pr(A_1) + \Pr(A_2)$. This holds for any number n of mutually exclusive events, so that $\Pr(A_1 \cup A_2 \ldots \cup A_n) = \sum_{i=1}^{n} \Pr(A_i)$.[5]

These three principles—collectively known as the Axioms of Probability—are the foundations of probability theory. They imply two useful properties of probability. First, the probability of an event or its complement must be 1.

$$\Pr(A \cup A^c) = \Pr(A) + \Pr(A^c) = 1 \tag{1.1}$$

Second, the probability of the union of any two sets can be decomposed into:

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B) \tag{1.2}$$

This result demonstrates that the total probability of two events is the probability of each event minus the probability of the event defined by the intersection. The elements in the intersection are in both sets, therefore subtracting this term avoids double counting.

Conditional Probability

Probability is commonly used to examine events in subsets of the full event space. Specifically, we are often interested in the probability of an event happening only if another event happens first. This concept is called *conditional probability* because we are computing a probability on that condition that another event occurs.

For example, we might want to determine the probability that a large financial institution fails given that another large financial institution has also failed. Note that the probability of a large financial institution failing is normally quite low. However, when one large financial institution fails (e.g., Lehmann Brothers in 2008), the probability of another failure is likely to be higher. This difference is important, especially in risk management.

Conditional probability can also be used to incorporate additional information to update unconditional probabilities. For example, consider two randomly chosen students preparing to take a risk management exam, which is passed (on average) by 50% of test takers. Based on only this information, one would estimate that both Student X and Student Y have a 50% chance of passing the exam.

Now suppose that the average length of study time needed to pass the exam is 200 hours. Suppose further that Student X studied for less than 100 hours, whereas Student Y studied for more than 400 hours. Common sense would say that Student Y has a higher probability of passing compared to Student X. In fact, a survey might find that the probability of passing for students who studied less than 100 hours is 10%, whereas the probability of passing for students who studied more than 400 hours is 80%. Using this new information, the conditional probability that X passes the exam is 10%, while the conditional probability that Y passes is 80%.

The conditional probability of the event A occurring given that B has occurred is denoted as $\Pr(A|B)$ and can be calculated as:

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} \tag{1.3}$$

The probability of observing an event in A, given that an event in B has been observed, depends on the probability of observing events that are in both rescaled by the probability that an event in B occurs. In other words, conditional probability operates as if B is the event space and A is an event inside this new, restricted space.

As a simple example, note that the probability of rolling a three on a fair die is 1/6. This is because the event of rolling a three occurs in one outcome in a set of six possible outcomes: {1, 2, 3, 4, 5, 6}. However, if it is known that the number rolled is odd, then the probability of rolling a three is 1/3 because this additional information restricts the set of outcomes to three possibilities: {1, 3, 5}.

Figure 1.2 demonstrates this idea using sets. The left panel shows the unconditional probability where Ω is the event space and A and B are events. The right panel shows the conditional probability of A given B. The numeric values show the probability of each region, meaning that the unconditional probability of A is $\Pr(A) = 16\%$ and the conditional probability of A given B is $\Pr(A|B) = 12\%/48\% = 25\%$.

An important application of conditional probability is the Law of Total Probability. This law states that the total probability of an event can be reconstructed using conditional probabilities under the condition that the probability of the sets being conditioned is equal to 1.

[5] This result holds more generally for infinite collections of mutually exclusive events under some technical conditions.

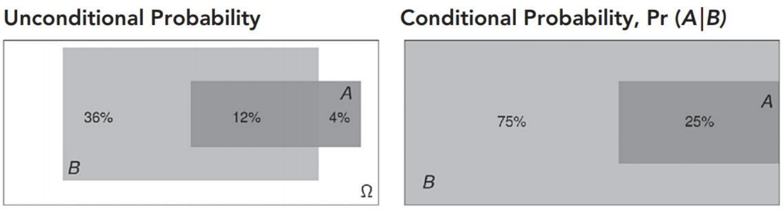


Figure 1.2 The left panel shows the unconditional probability of *A* and *B*. The right panel shows the conditional probability of *A* given *B*. The numeric values indicate the probability of each region.

For example, $\Pr(B) + \Pr(B^c) = 1$, and so:

$$\Pr(A) = \Pr(A|B)\Pr(B) + \Pr(A|B^c)\Pr(B^c) \quad (1.4)$$

This property holds for any collection of mutually exclusive events $\{B_i\}$, where $\sum \Pr(B_i) = 1$. Intuitively, this property must hold because each outcome in *A* must occur in one (and only one) of the B_i.

Figure 1.3 illustrates the Law of Total Probability for events *A*, B_1, B_2, B_3, and B_4. Note that the probability that *A* occurs is 9% + 12% + 14% = 35%. The four conditional probabilities corresponding to each set B_i are

$$\Pr(A|B_1) = 14\%/30\% = 46.6\%,$$

$$\Pr(A|B_2) = 9\%/22\% = 40.9\%,$$

$$\Pr(A|B_3) = 12\%/28\% = 42.8\%,$$

$$\Pr(A|B_4) = 0$$

These conditional probabilities are then rescaled by the probabilities of the conditioning event to recover the total probability.

$$\sum_{i=1}^{4} \Pr(A|B_i)\Pr(B_i) = 46.6\% \times 30\% + 40.9\% \times 22\% + 42.8\% \times 28\% + 0 \times 20\% = 35\%.$$

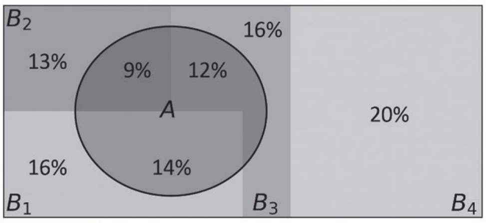


Figure 1.3 The four events B_1 through B_4 are mutually exclusive. *A* is the event indicated by the dark circle, and $Pr(A) = 35\%$. The probability of each event B_i includes the values that overlap with *A* as well as those that do not. For example, $Pr(B_1) = 16\% + 14\% = 30\%$.

Example: SIFI Failures

Systemically Important Financial Institutions (SIFIs) are a category of designated large financial institutions that are deeply connected to large portions of the economy. They are usually banks, although other types of institutions such as insurers, asset managers, or central counterparties may be designated as SIFIs as well. SIFIs are subject to additional regulation and supervision, as the failure of even one of them could lead to a major disruption in financial markets.

Suppose that the chance of at least one SIFI failing in any given year is 1%. Suppose further that when at least one SIFI fails, there is a 20% chance that the number of failing SIFIs is exactly 1, 2, 3, 4 or 5. In other words, there is a 0.2% chance that one SIFI fails, 0.2% chance that two institutions fail, and so on. Finally, assume that more than five SIFIs cannot fail, because governments and central banks would intervene.

If we define the event that one or more SIFIs fail as E_1, then $\Pr(E_1) = 1\%$. This is a small probability. However, if we are interested in the probability of two or more failures given that at least one has occurred, then this is

$$\Pr(E_2|E_1) = \frac{\Pr(E_1 \cap E_2)}{\Pr(E_1)} = \frac{\Pr(E_2)}{\Pr(E_1)} = \frac{0.8\%}{1\%} = 80\%$$

Here, E_2 is the event that two or more SIFIs fail, and it is a subset of E_1 by construction (because E_1 is the event of one or more failures). Because $E_1 \cap E_2 = E_2$, $\Pr(E_1 \cap E_2) = \Pr(E_2)$.

This conditional probability tells us that, while it would be surprising to see a SIFI failure in any given year, there is very likely to be two or more failures if there is at least one failure. The left panel of Figure 1.4 graphically illustrates this point.

1.2 INDEPENDENCE

Independence is a core concept that is exploited throughout statistics and econometrics. Two events are independent if the

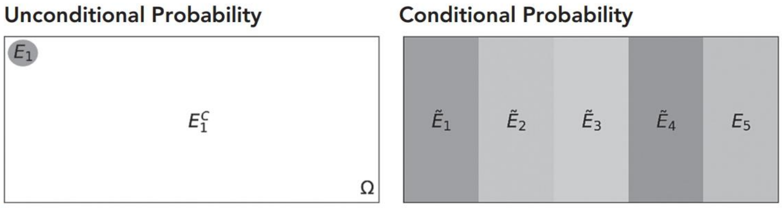


Figure 1.4 **The left panel illustrates the unconditional probability of at least one SIFI failing. The right panel illustrates the conditional probability of 1, 2, 3, 4 or 5 failures given that at least one SIFI has failed. The $\tilde{E}_i$ are the events that exactly *i* SIFIs fail. These differ from E_i, which measure the event that *i* or more institutions fail. However, $\tilde{E}_5 = E_5$ because the maximum number of failures is assumed to be five.**

probability that one event occurs does not depend on whether the other event occurs. When the two events *A* and *B* are independent, then:

$$\Pr(A \cap B) = \Pr(A)\Pr(B) \tag{1.5}$$

This implies that for independent events, the conditional probability of the event is equal to the unconditional probability of the event. When *A* and *B* are independent, then:

$$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)} = \frac{\Pr(A)\Pr(B)}{\Pr(A)} = \Pr(B) \tag{1.6}$$

It is also the case that $\Pr(A|B) = \Pr(A)$.

If *A* and *B* are mutually exclusive, then *B* cannot occur if *A* occurs. This means that the outcome of *A* affects $\Pr(B)$ and so mutually exclusive events are not independent. Another way to see this is to note that when $\Pr(A)$ and $\Pr(B)$ are both greater than 0, then $\Pr(A \cap B) = \Pr(A) \times \Pr(B) > 0$ (i.e., there must be a positive probability that both occur). However, mutually exclusive events have $\Pr(A \cap B) = 0$ and thus cannot be independent.

Independence can be generalized to any number of events using the same principle: the joint probability of the events is the product of the probability of each event. In other words, if $A_1, A_2, \ldots, A_n$ are independent events, then $\Pr(A_1 \cap A_2 \cap \ldots \cap A_n) = \Pr(A_1) \times \Pr(A_2) \times \ldots \times \Pr(A_n)$.

Conditional Independence

Like probability, independence can be redefined to hold conditional on another event. Two events *A* and *B* are conditionally independent if:

$$\Pr(A \cap B|C) = \Pr(A|C)\Pr(B|C) \tag{1.7}$$

The conditional probability $\Pr(A \cap B|C)$ is the probability of an outcome that is in both *A* and *B* occurring given that an outcome in *C* occurs.

Note that two types of independence—unconditional and conditional—do not imply each other. Events can be both unconditionally dependent (i.e., not independent) and conditionally independent. Similarly, events can be unconditionally independent, yet conditional on another event they may be dependent.

The left-hand panel of Figure 1.5 contains an illustration of two independent events. Each of the events *A* and *B* has a 40% chance of occurring. The intersection has the probability

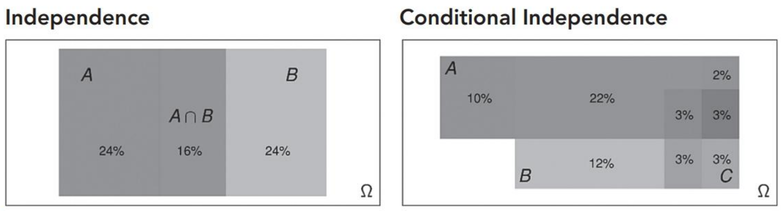


Figure 1.5 **The left panel shows two events are independent so that the joint probability, $\Pr(A \cap B)$, is the same as the product of the probability of each event, $\Pr(A) \times \Pr(B)$. The right panel contains an example of two events, *A* and *B*, which are dependent but also conditionally independent given *C*.**

$Pr(A) \times Pr(B)$ ($= 40\% \times 40\% = 16\%$) occurring, and so the requirements for independence are satisfied.

Meanwhile, the right-hand panel shows an example where A and B are unconditionally dependent: $Pr(A) = 40\%$, $Pr(B) = 40\%$ and $Pr(A \cap B) = 25\%$ (rather than 16% as would be required for independence). However, conditioning on the event C restricts the space so that $Pr(A|C) = 50\%$, $Pr(B|C) = 50\%$ and $Pr(A|C)Pr(B|C) = 25\%$, meaning that these events are conditionally independent. A further implication of conditional independence is that $Pr(A|B \cap C) = Pr(A|C) = 50\%$ and $Pr(B|A \cap C) = Pr(B|C) = 50\%$, which both hold.

1.3 BAYES' RULE

Bayes' rule provides a method to construct conditional probabilities using other probability measures. It is both a simple application of the definition of conditional probability and an extremely important tool. The formal statement of Bayes' rule is

$$Pr(B|A) = \frac{Pr(A|B)\,Pr(B)}{Pr(A)} \qquad (1.8)$$

This is a simple rewriting of the definition of conditional probability, because $Pr(A|B) = Pr(A \cap B)/Pr(B)$ so that $Pr(A \cap B) = Pr(A|B)\,Pr(B)$. Reversing these arguments also gives $Pr(A \cap B) = Pr(B|A)\,Pr(A)$. Equating these two expressions and solving for $Pr(B|A)$ produces Bayes' rule. This approach uses the unconditional probabilities of A and B, as well as information about the conditional probability of A given B, to compute the probability of B given A.

Bayes' rule has many financial and risk management applications. For example, suppose that 10% of fund managers are superstars. Superstars have a 20% chance of beating their benchmark by more than 5% each year, whereas normal fund managers have only a 5% chance of beating their benchmark by more than 5%.

Suppose there is a fund manager that beat her benchmark by 7%. What is the chance that she is a superstar? Here, event A is the manager significantly outperforming the fund's benchmark. Event B is that the manager is a superstar. Using Bayes' rule:

$$Pr(\text{Star}|\text{High Return}) = \frac{Pr(\text{High Return}|\text{Star})\,Pr(\text{Star})}{Pr(\text{High Return})}$$

The probability of a superstar is 10%. The probability of a high return if she is a superstar is 20%. The probability of a high return from either type of manager is

$$\begin{aligned} Pr(\text{High Return}) &= Pr(\text{High Return}|\text{Star})\,Pr(\text{Star}) \\ &\quad + Pr(\text{High Return}|\text{Normal})\,Pr(\text{Normal}) \\ &= 20\% \times 10\% + 5\% \times 90\% = 6.5\% \end{aligned}$$

Combining these produces the conditional probability that the manager is a star given that she had a high return:

$$Pr(\text{Star}|\text{High Return}) = \frac{20\% \times 10\%}{6.5\%} = \frac{2\%}{6.5\%} = 30.8\%$$

In this case, a single high return updates the probability that a manager is a star from 10% (i.e., the unconditional probability) to 30.8% (i.e., the conditional probability). This idea can be extended to classifying a manager given multiple returns from a fund.

1.4 SUMMARY

Probability is essential for statistics, econometrics, and data analysis. This chapter introduces probability as a simple way to understand events that occur with different frequencies. Understanding probability begins with the sample space, which defines the range of possible outcomes and events that are collections of these outcomes. Probability is assigned to events, and these probabilities have three key properties.

1. Any event A in the event space $\mathcal{F}$ has $Pr(A) \geq 0$.
2. The probability of all events in Ω is 1 and therefore $Pr(\Omega) = 1$.
3. If the events A_1 and A_2 are mutually exclusive, then $Pr(A_1 \cup A_2) = Pr(A_1) + Pr(A_2)$. This holds for any number of mutually exclusive events, so that $Pr(A_1 \cup A_2 \ldots \cup A_n) = \sum_{i=1}^{n} Pr(A_i)$.

These properties allow a complete description of event probabilities using a standard set of operators (e.g., union, intersection, and complement). Conditional probability extends the definition of probability to subsets of all events, so that a conditional probability is like an unconditional probability, but only within a smaller event space.

This chapter also examines independence and conditional independence. Two events are independent if the probability of both events occurring is the product of the probability of each event. In other words, events are independent if there is no information about the likelihood of one event given knowledge of the other. Independence extends to conditional independence, which replaces the unconditional probability with a conditional probability. An important feature of conditional independence is that events that are not unconditionally independent may be conditionally independent.

Finally, Bayes' rule uses basic ideas from probability to develop a precise expression for how information about one event can be used to learn about another event. This is an important practical tool because analysts are often interested in updating their beliefs using observed data.

The following questions are intended to help candidates understand the material. They are *not* actual FRM exam questions.

QUESTIONS

Short Concept Questions

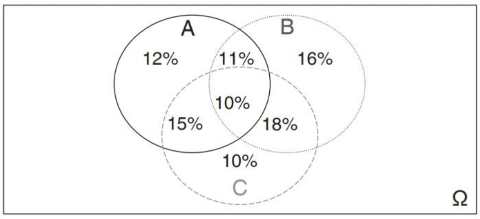
1.1 Can independent events be mutually exclusive?

1.2 Can Bayes' rule be helpful if *A* and *B* are independent? What if *A* and *B* are perfectly dependent so that *B* is a subset of *A*?

1.3 What are the sample spaces, events, and event spaces if you are interested in measuring the probability of a corporate takeover and whether it was hostile or friendly?

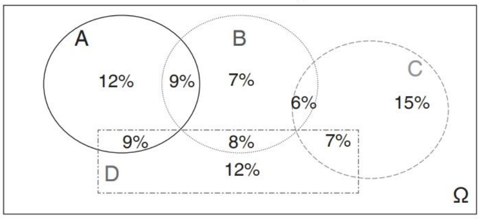
Practice Questions

1.4 Based on the percentage probabilities in the diagram below, what are the values of the following?



- **a.** $\Pr(A)$
- **b.** $\Pr(A|B)$
- **c.** $\Pr(B|A)$
- **d.** $\Pr(A \cap B \cap C)$
- **e.** $\Pr(B|A \cap C)$
- **f.** $\Pr(A \cap B|C)$
- **g.** $\Pr(A \cup B|C)$
- **h.** Considering the three distinct pairs of events, are any of these pairs independent?

1.5 Based on the percentage probabilities in the plot below, what are the values of the following?



- **a.** $\Pr(A^C)$
- **b.** $\Pr(D|A \cup B \cup C)$
- **c.** $\Pr(A|A)$
- **d.** $\Pr(B|A)$
- **e.** $\Pr(C|A)$
- **f.** $\Pr(D|A)$
- **g.** $\Pr((A \cup D)^C)$
- **h.** $\Pr(A^C \cap D^C)$
- **i.** Are any of the four events pairwise independent?

1.6 Continue the application of Bayes' rule to compute the probability that a manager is a star after observing two years of "high" returns.

1.7 There are two companies in an economy, Company A and Company B. The default rate for Company A is 10%, and the default rate for Company B is 20%. Assume that defaults for the two companies occur independently.

- **a.** What is the probability that both companies default?
- **b.** What is the probability that either Company A or Company B defaults?

1.8 Credit card companies rapidly assess transactions for fraud. In each day, a large card issuer assesses 10,000,000 transactions. Of these, 0.001% are fraudulent. If their algorithm identifies 90% of all fraudulent transactions but also 0.0001% of legitimate transactions, what is the probability that a transaction is fraudulent if it has been flagged?

1.9 If fund managers beat their benchmark at random (meaning a 50% chance that they beat their benchmark), and annual returns are independent, what is the chance that a fund manager beats her benchmark for ten years in a row?

The following questions are intended to help candidates understand the material. They are *not* actual FRM exam questions.

ANSWERS

Short Concept Questions

1.1 No. For independent events, the probability of observing both must be the product of the individual probabilities, and mutually exclusive events have probability 0 of simultaneously occurring.

1.2 Bayes' rule says that $\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$. If these events are independent, then $\Pr(B|A) = \Pr(B)$ so that $\Pr(A|B) = \Pr(A)$. B has no information about A and so updating with Bayes' rule never changes the conditional probability. If these events are perfectly dependent, then $\Pr(B|A) = 1$, so that $\Pr(A|B)$ is just the ratio of the probabilities, $\Pr(A)/\Pr(B)$. Here the probability also only depends on unconditional probabilities and so never changes.

1.3 Using the notation No Takeover (NT), Hostile Takeover (HT) and Friendly Takeover (FT), the sample space is {*NT*, *HT*, *HT*}. The events are all distinct combinations of the sample space including the empty set, ϕ, *NT*, *HT*, *FT*, {*NT*, *HT*}, {*NT*, *FT*}, {*HT*, *FT*}, {*NT*, *HT*, *FT*}. The event space is the set of all events.

Solved Problems

1.4 **a.** $\Pr(A) = 12\% + 11\% + 15\% + 10\% = 48\%$

b. $\Pr(A|B) = \Pr(A\cap B)/\Pr(B) = (11\% + 10\%)/(11\% + 10\% + 16\% + 18\%) = 38.2\%$

c. $\Pr(B|A) = \Pr(A\cap B)/\Pr(A) = (11\% + 10\%)/48\% = 43.8\%$

d. $\Pr(A\cap B\cap C) = 10\%$

e. $\Pr(B|A\cap C) = \Pr(B\cap A\cap C)/\Pr(A\cap C) = 10\%/(15\% + 10\%) = 40\%$

f. $\Pr(A\cap B|C) = \Pr(A\cap B\cap C)/\Pr(C) = 10\%/(15\% + 10\% + 18\% + 10\%) = 18.9\%$

g. $\Pr(A\cup B|C) = \Pr((A\cup B)\cap C)/\Pr(C) = (15\% + 10\% + 18\%)/53\% = 81.1\%$

h. We can use the rule that events are independent if their joint probability is the product of the probability of each. $\Pr(A) = 48\%$, $\Pr(B) = 55\%$, $\Pr(C) = 53\%$, $\Pr(A\cap B) = 21\% \neq (48\% \times 55\%)$, $\Pr(A\cap C) = 25\% \neq (48\% \times 53\%)$, $\Pr(B\cap C) = 28\% \neq (55\% \times 53\%)$. None of these events are pairwise independent.

1.5 **a.** $1 - \Pr(A) = 100\% - 30\% = 70\%$

b. This value is $\Pr(D\cap(A\cup B\cup C))/\Pr(A\cup B\cup C)$. The total probability in the three areas A, B, and C is 73%. The overlap of D with these three is $9\% + 8\% + 7\% = 24\%$, and so the conditional probability is $\frac{24\%}{73\%} = 33\%$.

c. This is trivially 100%.

d. $\Pr(B\cap A) = 9\%$. The conditional probability is $\frac{9\%}{30\%} = 30\%$.

e. There is no overlap and so $\Pr(C\cap A) = 0$.

f. $\Pr(D\cap A) = 9\%$. The conditional probability is 30%.

g. This is the total probability not in A or D. It is $1 - \Pr(A\cup D) = 1 - (\Pr(A) + \Pr(D) - \Pr(A\cap D)) = 100\% - (30\% + 36\% - 9\%) = 43\%$.

h. This area is the intersection of the space not in A with the space not in D. This area is the same as the area that is not in A or D, $\Pr((A\cup D)^C)$ and so 43%.

i. The four regions have probabilities $A = 30\%$, $B = 30\%$, $C = 28\%$ and $D = 36\%$. The only region that satisfied the requirement that the joint probability is the product of the individual probabilities is A and B because $\Pr(A\cap B) = 9\% = \Pr(A)\Pr(B) = 30\% \times 30\%$.

1.6 Consider the three scenarios: (High, High), (High, Low) and (Low, Low).

We are interested in Pr (*Star*|*High, High*) using Bayes' rule, this is equal to

$$\frac{\Pr(High, High|Star)\Pr(Star)}{\Pr(High, High)}.$$

Stars produce high returns in 20% of years, and so $\Pr(High, High|Star) = 20\% \times 20\%$ Pr (*Star*) is still 10%. Finally, we need to compute Pr (*High, High*), which is

$\Pr(High, High|Star)\Pr(Star) + \Pr(High, High|Normal)$ $\Pr(Normal)$. This value is $20\% \times 20\% \times 10\% + 5\% \times 5\% = 90\% = 0.625\%$. Combining these values,

$$\frac{20\% \times 20\% \times 10\%}{0.625\%} = 64\%.$$

This is a large increase from the 30% chance after one year.

1.7 **a.** $0.10 \times 0.20 = 0.02 \rightarrow 2\%$

b. Calculate this in two ways:

i. Using Equation 1.1:

$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$

$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$

$= 10\% + 20\% - 2\% = 28\%.$

ii. Using the identity: $\Pr(A \cup A^C) = \Pr(A) + \Pr(A^C) = 1$

Let C be the event of no defaults. Then

$\Pr(C) = (1 - (1 - \Pr(A)) \times (1 - \Pr(B))$

$= (1 - 10\%) \times (1 - 20\%) = 0.9 \times 0.8 = 72\%.$

Then the complement of C is the event of any defaults $\Pr(C^C) = 1 - \Pr(C) = 28\%$.

1.8 We are interested in Pr(Fraud|Flag). This value is Pr(Fraud ∩ Flag)/Pr(Flag). The probability that a transaction is flagged is $0.001\% \times 90\% + 99.999\% \times 0.0001\% = 0.000999\%$. The Pr(Fraud ∩ Flag) $= 0.001\% \times 90\% = 0.0009\%$. Combining these values, $\frac{0.0009\%}{0.000999\%} = 90\%$. This indicates that 10% of the flagged transactions are not actually fraudulent.

1.9 If each year was independent, then the probability of beating the benchmark ten years in a row is the product of the probabilities of beating the benchmark each year, $\left(\frac{1}{2}\right)^{10} = \frac{1}{1{,}024} \approx 0.1\%$.

Random Variables 2

Learning Objectives

After completing this reading you should be able to:

- Describe and distinguish a probability mass function from a cumulative distribution function, and explain the relationship between these two.
- Understand and apply the concept of a mathematical expectation of a random variable.
- Describe the four common population moments.
- Explain the differences between a probability mass function and a probability density function.
- Characterize the quantile function and quantile-based estimators.
- Explain the effect of a linear transformation of a random variable on the mean, variance, standard deviation, skewness, kurtosis, median, and interquartile range.

Probabilities can be used to describe any situation with an element of uncertainty. However, random variables restrict attention to uncertain phenomena that can be described with numeric values. This restriction allows standard mathematical tools to be applied to the analysis of random phenomena. For example, the return on a portfolio of stocks is numeric, and a random variable can therefore be used to describe its uncertainty. The default on a corporate bond can be similarly described using numeric values by assigning one to the event that the bond is in default and zero to the event that the bond is in good standing.

This chapter begins by defining a random variable and relating its definition to the concepts presented in the previous chapter. The initial focus is on discrete random variables, which take on distinct values. Note that most results from discrete random variables can be computed using only a weighted sum.

Two functions are commonly used to describe the chance of observing various values from a random variable: the probability mass function (PMF) and the cumulative distribution function (CDF). These functions are closely related, and each can be derived from the other. The PMF is particularly useful when defining the expected value of a random variable, which is a weighted average that depends on both the outcomes of the random variable and the probabilities associated with each outcome.

Moments are used to summarize the key features of random variables. A moment is the expected value of a carefully chosen function designed to measure a characteristic of a random variable. Four moments are commonly used in finance and risk management: the mean (which measures the average value of the random variable), the variance (which measures the spread/dispersion), the skewness (which measures asymmetry), and the kurtosis (which measures the chance of observing a large deviation from the mean).[1]

Another set of important measures for random variables includes those that depend on the quantile function, which is the inverse of the CDF. The quantile function, which can be used to map a random variable's probability to its realization (i.e., the actual value that subsequently occurs), defines two moment-like measures: the median (which measures the central tendency of a random variable) and the interquartile range (which is an alternative measure of spread). The quantile function is also used to simulate data from a random variable with a particular distribution later in this book.

The chapter concludes by examining continuous random variables, which produce values that lie in a continuous range.

[1] The skewness and kurtosis are technically standardized versions of the third and fourth moment.

Whereas a discrete random variable depends on a probability *mass* function, a continuous random variables depends on a probability *density* function (PDF). However, this difference is of little consequence in practice, and the concepts introduced for discrete random variables all extend to continuous random variables.

2.1 DEFINITION OF A RANDOM VARIABLE

The axioms of probability are general enough to describe many forms of randomness (e.g., a coin flip, a draw from a well-shuffled deck of cards, or a future return on the S&P 500). However, directly applying probability can be difficult because it is defined on events, which are abstract concepts. Random variables limit attention to random phenomena which can be described using numeric values. This covers a wide range of applications relevant to risk management (e.g., describing asset returns, measuring defaults, or quantifying uncertainty in parameter estimates). The restriction that random variables are only defined on numeric values simplifies the mathematical tools used to characterize their behavior.

A random variable is a function of ω (i.e., an event in the sample space Ω) that returns a number x. It is conventional to write a random variable using upper-case letters (e.g., X, Y, or Z) and to express the realization of that random variable with lower-case letters (e.g., x, y, or z). It is important to distinguish between these two: A random variable is a function, whereas a realization is a number. The relationship between a realization and its corresponding random variable can be succinctly expressed as:

$$x = X(\omega) \tag{2.1}$$

For example, let X be the random variable defined by the roll of a fair die. Then we can denote x as the result of a single roll (i.e., one event). The probability that the random variable is equal to five can be expressed as:

$$\Pr(X = x) \text{ when } x = 5$$

Univariate random variables are frequently distinguished based on the range of values produced. The two classes of random variables are discrete and continuous. A discrete random variable is one that produces distinct values.[2] A continuous random variable produces values from an uncountable set, (e.g., any number on the real line $\mathbb{R}$).

[2] The set of values that can be produced by a discrete random variable can be either finite or have an infinite range. When the set of values is infinite, the set must be countable.

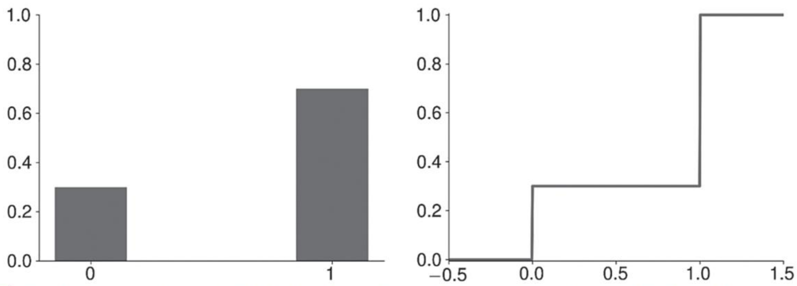


Figure 2.1 **The left panel shows the PMF of a Bernoulli random variable with $p = 0.7$. The right panel shows the corresponding CDF.**

2.2 DISCRETE RANDOM VARIABLES

A discrete random variable assigns a probability to a set of distinct values. This set can be either finite or contain a countably infinite set of values.[3] An important example of a discrete random variable with a finite number of values is a Bernoulli random variable, which can only take a value of 0 or 1. Bernoulli random variables are frequently encountered when measuring binary random events (e.g., the default of a loan).

Because the values of random variables are always numerical, they can be described precisely using mathematical functions. The set of values that the random variable may take is called the support of the function. For example, the support for a Bernoulli random variable is {0,1}. Two functions are frequently used when describing the properties of a discrete random variable's distribution.

The first function is known as the probability mass function (PMF), which returns the probability that a random variable takes a certain value. Because the PMF returns probabilities, any PMF must have two properties:

1. The value returned from a PMF must be non-negative.
2. The sum across all values in the support of a random variable must be one.

For example, if X is a Bernoulli random variable and the probability that $X = 1$ is denoted by p (where p is between 0 and 1), the PMF of X is:

$$f_X(x) = p^x(1 - p)^{1-x} \tag{2.2}$$

and the inputs to the PMF are either 0 or 1.

[3] The leading example of a countably infinite set is the collection of all integers, and any countably infinite set has a one-to-one relationship to the integers.

Note that $f_X(0) = p^0(1 - p)^1 = 1 - p$ and $f_X(1) = p^1(1 - p)^0 = p$, so that the probability of observing 0 is $1 - p$ and the probability of observing 1 is p. The PMF is denoted by $f_X(x)$ to distinguish the random variable X underlying the mass function from the realization of the random variable (i.e., x).

The PMF of a Bernoulli random variable can be equivalently expressed using a list of values:

$$f_X(x) = \begin{cases} 1 - p & x = 0 \\ p & x = 1 \end{cases} \tag{2.3}$$

The counterpart of a PMF is the cumulative distribution function (CDF), which measures the total probability of observing a value less than or equal to the input x (i.e., $\Pr(X \le x)$). In the case of a Bernoulli random variable, the CDF is a simple step function:

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1 - p & 0 \le x < 1 \\ 1 & x \ge 1 \end{cases} \tag{2.4}$$

Because the CDF measures the total probability that $X \le x$, it must be monotonic and increasing in x. Note that while the PMF is a discrete function (i.e., it only has support for 0 and 1), the CDF returns $\Pr(X \le x)$ and is therefore a continuous function that has support for any value of x, not just 0 and 1.

Figure 2.1 shows the PMF (left panel) and CDF (right panel) of a Bernoulli random variable with parameter $p = 0.7$ (i.e., there is a 70% chance of observing $x = 1$). The PMF is only defined for the two possible outcomes (i.e., 0 or 1). In contrast, the CDF is defined for all values of x, and so is a step function that has the value 0 for $x < 0$, 0.3 for $0 \le x < 1$, and 1 for $x \ge 1$.

Note that the PMF is denoted by $f_X(x)$, whereas the CDF is denoted by $F_X(x)$. The PMF and the CDF are closely related, and

the CDF can always be expressed as the sum of the PMF for all values in the support that are less than or equal to x:

$$F_X(x) = \sum_{t \in R(X), t \leq x} f_X(t) \tag{2.5}$$

where $R(X)$ is used to denote the range of realizations that may be produced from the random variable X and $\in$ means "an element of." Conversely, the PMF is the difference of the CDFs evaluated at consecutive values in the support of X. For a Bernoulli random variable:

$$f_X(x) = F_X(x) - F_X(x-1) \tag{2.6}$$

for $x \in \{0,1\}$.

2.3 EXPECTATIONS

The expected value of a random variable is denoted $E[X]$ and is defined as:

$$E[X] = \sum_{x \in R(X)} x \Pr(X = x), \tag{2.7}$$

where $R(X)$ is the range of values of the random variable X. Thus, an expectation is simply an average of the values in the support of X weighted by the probability that a value x is observed. For example, when X is a Bernoulli random variable where the probability of observing 1 is p:

$$E[X] = 0 \times (1 - p) + 1 \times p = p \tag{2.8}$$

Functions of random variables are also random variables and therefore have expected values. The expectation of a function f of a random variable X is defined as:

$$E[f(X)] = \sum_{x \in R(X)} f(x) Pr(X = x), \tag{2.9}$$

where $f(x)$ is a function of the realization value x.

For example, when X is a Bernoulli random variable, then the expected value of the exponential of X is

$$E[\exp(X)] = \exp(0) \times (1 - p) + \exp(1) \times p$$
$$= (1 - p) + p \exp(1)$$

As an example, consider a call option (i.e., a derivative contract with a payoff that is a nonlinear function of the future price of the underlying asset). The payoff of the call option (at the expiry date) is $c(S) = \max(S - K, 0)$, where S is the value of the underlying asset at expiry and K is the strike price. If the asset price is above the strike price, the call option pays out $S - K$. If the price is below the strike price, the call option pays zero. Valuing a call option involves computing the expectation of a function (i.e., the payoff) of the price (i.e., a random variable).

Suppose the value of an asset S will take on one of three values (i.e., 20, 50, or 100) in one year with probabilities 0.2, 0.5, and 0.3, respectively. The PMF of S is

$$f_S(s) = \begin{cases} 0.2 & \text{if } s = 20 \\ 0.5 & \text{if } s = 50 \\ 0.3 & \text{if } s = 100 \end{cases}$$

If the strike price $K = 40$, then the payoff of the call option can be written as

$$c(S) = \max(S - 40, 0)$$

Thus, the expected payoff of the call option is

$$E[c(S)] = 0.2 \times \max(20 - 40,0) + 0.5 \times \max(50 - 40,0) + 0.3 \times \max(100 - 40,0)$$
$$= 0.2 \times 0 + 0.5 \times 10 + 0.3 \times 60$$
$$= 23.$$

Properties of the Expectation Operator

The expectation operator $E[\cdot]$ takes a random variable and computes a weighted average of its possible values. $E[\cdot]$ has a particularly useful property in that it is a linear operator. For example, if a, b, and c are constants, and X and Y are random variables, then $E[a + bX + cY] = a + bE[X] + cE[Y]$.

An immediate implication of linearity is that the expectation of a constant is the constant (e.g., $E[a] = a$). A related implication is that the expectation of the expectation is just the expectation (e.g., $E[E[X]] = E[X]$). Furthermore, the expected value of a random variable is a constant (and not a random variable).

While linearity is a useful property, the expectation operator does *not* pass through nonlinear functions of X. For example, $E[1/X] \neq 1/E[X]$ whenever X is a non-degenerate random variable.[4] This property holds for other nonlinear functions so that in general[5] $E[g(X)] \neq g(E[X])$.

2.4 MOMENTS

As stated previously, moments are a set of commonly used descriptive measures that characterize important features of random variables. Specifically, a moment is the expected value of a function of a random variable. The first moment is defined as the expected value of X:

$$\mu_1 = E[X] \tag{2.10}$$

[4] A degenerate random variable is a one that assigns all of the probability mass to a single point. Hence it always returns the same value, and is thus a constant.

[5] Except when $g(X)$ is a linear function.

JENSEN'S INEQUALITY

Jensen's inequality is a useful rule that provides some insight into the expectation of a nonlinear function of a random variable. Jensen's inequality applies when a nonlinear function is either concave or convex.

- A function $g(x)$ is concave if $g((1 - t)x + tx) \geq (1 - t)g(x) + tg(x)$, for all t in the interval [0,1] and x in the domain of g.
- A function $h(x)$ is convex if $h((1 - t)x + tx) \leq (1 - t)h(x) + th(x)$, for all t in the interval [0,1] and x in the domain of h.

If $h(X)$ is a convex function of X, then $E[h(X)] > h(E[X])$. Similarly, if $g(X)$ is a concave function, then $E[g(x)] < g(E[X])$.

Convex and concave functions are common in finance (e.g., many derivative payoffs are convex in the price of the underlying asset). It is useful to understand that the expected value of a convex function is larger than the function of the expected value.

Figure 2.2 illustrates Jensen's inequality for both convex and concave functions. In each panel, the random variable is discrete and can only take on one of two values, each with probability of 50%. The two points in the center of each graph show the function of the expectation (which is always on the curved graph of the function) and the expected value of the function (which is always on the straight line connecting the two values).

As an example of a convex function, suppose you play a game in which you roll a die and receive the square of the number rolled. Let X represent the die roll. We have that the payoff function is $f(X) = X^2$. Then:

$$E[X] = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = 3.5$$

$$E[f(X)] = E[X^2] = \frac{1 + 4 + 9 + 16 + 25 + 36}{6} = 15.167$$

and:

$$f(E[X]) = f(3.5) = 3.5^2 = 12.5.$$

Thus we can see that $E[f(x)] > f(E[X])$, as required for a convex function.

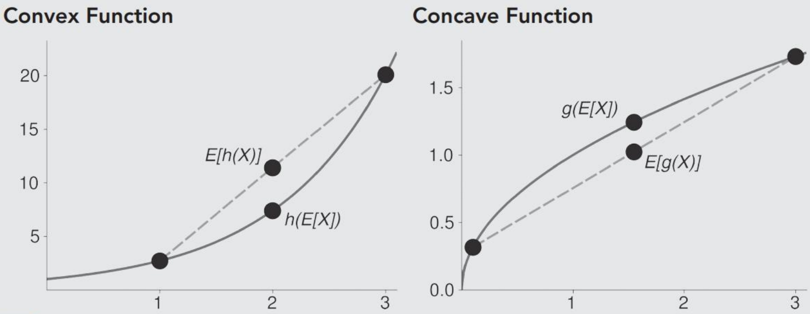


Figure 2.2 **These two panels demonstrate Jensen's inequality. The left shows the difference between the expected value of the function E[*h*(x)] and the function at the expected value, *h*(E[X]), for a convex function. The right shows this difference for a concave function.**

The second and higher moments all share the same form and only differ in one parameter. These moments are defined as:

$$\mu_r = E[(X - E[X])^r] \qquad (2.11)$$

where r denotes the order of the moment (2 for the second, 3 for the third, and so on). In the definition of the second and higher moments. Formally these moments are known as central moments because they are all centered on the first moment, $E[X]$.

There is also a second class of moments, called non-central moments.[6] These moments are not centered around the mean and thus are defined as:

$$\mu_r^{NC} = E[X^r] \qquad (2.12)$$

[6] The non-central moments, which do not subtract the first moment, are less useful for describing random variables because a change in the first moment changes all higher order non-central moments. They are, however, often simpler to compute for many random variables, and so it is common to compute central moments using non-central moments.

Non-central moments and central moments contain the same information and the r^{th} central moment can be directly constructed from the first r non-central moments. For example, the second central moment can be calculated as:[7]

$$E[(X - E[X])^2] = E[X^2] - E[X]^2 = \mu_2^{NC} - (\mu_1^{NC})^2$$

Note that the first central and non-central moments are the same:

$$\mu_1 = \mu_1^{NC}$$

The Four Named Moments

The first four moments are commonly used to describe random variables. They are the mean, the variance, the skewness, and the kurtosis.

The mean is the first moment and is denoted

$$\mu = E[X]$$

The mean is the average value of X and is also referred to as the location of the distribution.

The variance is the second moment and is denoted by:

$$\sigma^2 = E[(X - E[X])^2] = E[(X - \mu)^2]$$

The variance is a measure of the dispersion of a random variable. It measures the squared deviation from the mean, and thus it is not sensitive to the direction of the difference relative to the mean.

The standard deviation is denoted by σ and is defined as the square root of the variance (i.e., $\sqrt{\sigma^2}$). It is a commonly reported alternative to the variance, as it is a more natural measure of dispersion and is directly comparable to the mean (because the units of measurement of the standard deviation are the same as those of the mean).[8]

The final two named moments are both standardized versions of central moments. The skewness is defined as:

$$\text{skew}(X) = \frac{E[(X - E[X])^3]}{\sigma^3} = E\left[\left(\frac{X - \mu}{\sigma}\right)^3\right] \quad (2.13)$$

The final version in the definition clearly demonstrates that the skewness measures the cubed standardized deviation. Note that the random variable $(X - \mu)/\sigma$ is a standardized version of X that has mean 0 and variance 1.

This standardization makes the skewness comparable across distributions with different first and second moments and means that the skewness is unit-free by construction.

The skewness measures asymmetry in a distribution, because the third power depends on the sign of the difference. Negative values of the skewness indicate that the chance of observing a large (in magnitude) negative value is higher than the chance of observing a similarly large positive value. Asset returns are frequently found to have negative skewness since asset prices have a tendency to rise more often than they fall, but in smaller amounts.

The fourth standardized moment, known as the kurtosis, is defined as:

$$\text{kurtosis}(X) = \frac{E[(X - E[X])^4]}{\sigma^4} = E\left[\left(\frac{X - \mu}{\sigma}\right)^4\right] \quad (2.14)$$

Its definition is analogous to that of skewness, except that kurtosis uses the fourth power instead of the third. The use of the fourth power means the kurtosis is especially sensitive to large deviations of either sign, which makes it useful in analyzing the chance of observing a large return of either sign. Specifically, the kurtosis of a random variable is commonly benchmarked against that of a normally distributed random variable, which is 3. Random variables with kurtosis greater than 3, a common characteristic of financial return distributions, are described as being heavy-tailed/fat-tailed. Like skewness, kurtosis is naturally unit-free and can be directly compared across random variables with different means and variances.

STANDARDIZING RANDOM VARIABLES

When X has mean μ and variance σ^2, a standardized version of X can be constructed as $\frac{X - \mu}{\sigma}$. This variable has mean 0 and unit variance (and standard deviation). These can be verified because:

$$E\left[\frac{X - \mu}{\sigma}\right] = \frac{E[X] - \mu}{\sigma} = \frac{\mu - \mu}{\sigma} = 0$$

where μ and σ are constants and therefore pass through the expectation operator. The variance is

$$V\left[\frac{X - \mu}{\sigma}\right] = V\left[\frac{X}{\sigma} - \frac{\mu}{\sigma}\right] = V\left[\frac{X}{\sigma}\right] = \frac{1}{\sigma^2}V[X] = \frac{\sigma^2}{\sigma^2} = 1$$

because adding a constant has no effect on the variance (i.e., $V[X - c] = V[X]$) and the variance of bX is

$$b^2V[X]$$

when b and c are constants.

[7] Note that $E[(X - E[X])^2] = E[X^2 - 2xE[X] + E[X]^2] = E[X^2] - 2E[X]E[X] + E[X]^2 = E[X^2] - E[X]^2$.

[8] The units of the variance are the squared units of the random variable. For example, if X is the return on an asset, measured in percent, then the units of the mean and standard deviation are both percentages. The units of the variance are squared percentages.

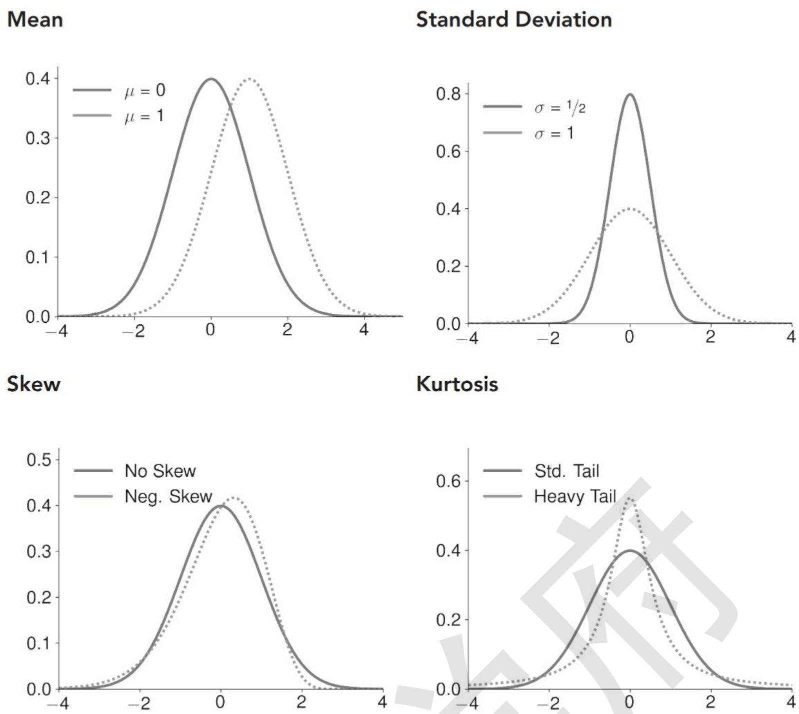


Figure 2.3 **The four panels demonstrate the effect of changes in the four moments. The top-left panel shows the effect of changing the mean. The top-right compares two distributions with the same mean but different standard deviations. The bottom-left panel shows the effect of negative skewness. The bottom-right shows the effect of heavy tails, which produces excess kurtosis.**

Figure 2.3 shows the effect of changing each moment. The top-left panel compares the distribution of two random variables that differ only in the mean. The top-right compares two random variables with the same mean but different standard deviations. In this case, the random variable with a smaller standard deviation has both a narrower and a taller density. Note that this increase in height is required when the scale is reduced to ensure that the total probability (as measured by the area under the curve) remains at one.

The bottom-left panel shows the effect of a negative skew on the distribution of a random variable. The skewed distribution has a short right tail and a longer left tail which extends to the edge of the graph. The bottom-right graph shows the effect of heavy tails, which produce excess kurtosis. The heavy-tailed distribution has more probability mass near the center and in the tails, and less in the intervals $(-2, -1)$ and $(1,2)$.

Moments and Linear Transformations

There are many situations in finance where it is convenient to rescale the data. For example, asset returns are commonly expressed as percentages for ease of interpretation. Another example is the situation where the mean of the original series is subtracted from every data point to achieve a variable with a mean of zero. Both of these are examples of linear transformations.

It is helpful to understand the effect of linear transformations on the first four moments of a random variable. As an example, consider

$$Y = a + bX,$$

where a and b are both constant values. It is common to refer to a as a location shift and b as a scale, because these directly affect the mean and standard deviation of Y, respectively. The mean of Y is:

$$E[Y] = a + bE[X].$$

This follows directly from the linearity of the expectation operator. The variance of Y is:

$$b^2V[X]$$

or equivalently:

$$b^2\sigma^2.$$

Note that the variance is unaffected by the location shift a because it only measures deviations around the mean. The standard deviation of Y is

$$\sqrt{b^2\sigma^2} = |b|\sigma.$$

The standard deviation is also insensitive to the shift by a but is linear in b. Finally, if b is positive (so that $Y = a + bX$ is an increasing transformation), then the skewness and kurtosis of Y are identical to the skewness and kurtosis of X. This is because both moments are defined on standardized quantities (e.g., $(Y - E[Y])/\sqrt{V[Y]}$), which removes the effect of the location shift by a and rescaling by b. If $b < 0$ (and thus $Y = a + bX$ is a decreasing transformation), then the skewness has the same magnitude but the opposite sign. This is because it uses an odd power (i.e., 3). The kurtosis, which uses an even power (i.e., 4), is unaffected when $b < 0$.

2.5 CONTINUOUS RANDOM VARIABLES

A continuous random variable is one that can take any of an infinite number of values. It can be more technically defined as a random variable with continuous support.

For example, it is common to use continuous random variables when modeling asset returns because they can take any value up to the accuracy level of the asset's price (e.g., dollars or cents). While many continuous random variables, such as yields or returns, have support for any value on the real number line (i.e., $\mathbb{R}$), others have support for a defined interval on the real number line (e.g., the interval [0,1] for the probability that it will rain tomorrow).

Continuous random variables use a probability density function (PDF) in place of the probability mass function. The PDF $f(X)$ returns a non-negative value for any input in the support of X. Note that a single value of $f(X)$ is technically not a probability because the probability of any single value is always 0. This is necessary because there are infinitely many values x in the support of X. Thus if the probability of each value was larger than 0, then the total probability would be larger than 1 and thus violate an axiom of probability.[9]

To see how this works, consider

- The interval [0,1], and
- A 10-sided die with the numbers {0.05,0.15, … ,0.95}.

Note that there is a 10% probability of getting any single number. Now suppose the die has 1,000 sides with the numbers {0.0005, … ,0.9995} and thus there is a 0.1% chance of any single number appearing. As the number of points (or number of sides on the die) increases, the probability of any specific value decreases. Therefore, as the number of points becomes infinite, the probability of a single point approaches zero. In other words, a continuous distribution is like a die with an infinite number of sides. In order to define a probability in a continuous distribution, it is therefore necessary to examine the values *between* points.

Recall that a PMF is required to sum to 1 across the support of X. This property also applies to a PDF, except that the sum is replaced by an integral:

$$\int_{-\infty}^{\infty} f_X(x)dx = 1$$

The PDF has a natural graphical interpretation and the area under the PDF between x_L and x_U is the probability that X will fall between these values, $\Pr(x_L < X < x_U)$. The shape of a PDF illustrates which ranges of values of X are more likely to occur. For example, the PDF for a normal distribution, with its familiar 'bell shape' is higher in the center, indicating that values of X near the mean of the distribution are more likely to occur than those further from the mean.

Meanwhile, the CDF of a continuous random variable is identical to that of a discrete random variable. It is a function that returns the total probability that the random variable X is less than or equal to the input (i.e., $\Pr(X \le x)$). For most common continuous random variables, the PDF and the CDF are directly related and the PDF can be derived from the CDF by taking the derivative:

$$f_X(x) = \frac{\partial F_X(x)}{\partial x}$$

Conversely, the CDF is the integral of the PDF up to x:

$$F_X(x) = \int_{-\infty}^{x} f_X(z)dz$$

[9] In fact, the total probability would be infinite because of the infinite number of possible values of X.

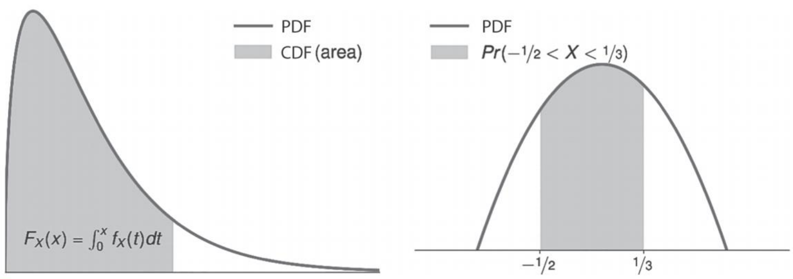


Figure 2.4 **The left panel shows the relationship between the PDF and CDF of a random variable. The right panel illustrates the probability of an interval.**

The left panel of Figure 2.4 illustrates the relationship between the PDF and CDF of a continuous random variable. The random variable has support on $(0,\infty)$, meaning that the CDF at the point x (i.e., $F_X(x)$) is the area under the curve between 0 and x.

The right panel shows the relationship between a PDF and the area in an interval. The probability $\Pr(-\frac{1}{2} < X \leq \frac{1}{3})$ is the area under the PDF between these two points. It can be computed from the CDF as:

$$F_X(1/3) - F_X(-1/2) = \Pr(X \leq 1/3) - \Pr(X \leq -1/2)$$

The probabilities $\Pr(a < X \leq b)$, $\Pr(a \leq X \leq b)$, $\Pr(a \leq X < b)$, and $\Pr(a < X < b)$ are the same when X is a continuous random variable since $\Pr(X = a) = 0$ and $\Pr(X = b) = 0$. However, this is not necessarily the case when X is a discrete random variable.

The expectation of a continuous random variable is also an integral. The mean E[X] is

$$\text{E}[X] = \int_{R(X)} x f_X(x) dx \qquad (2.15)$$

Superficially, this appears to be different from the expectation of a discrete random variable. However, an integral can be approximated as a sum so that:

$$\text{E}[X] \approx \Sigma x \Pr(X = x)$$

where the sum is over a range of values in the support of X. The relationship between continuous and discrete random variables is illustrated in Figure 2.5. The left panel shows a seven-point PMF that is a coarse approximation to the PDF of a normal random variable. This approximation has large errors (which are illustrated by gaps between the PMF and the PDF). The right

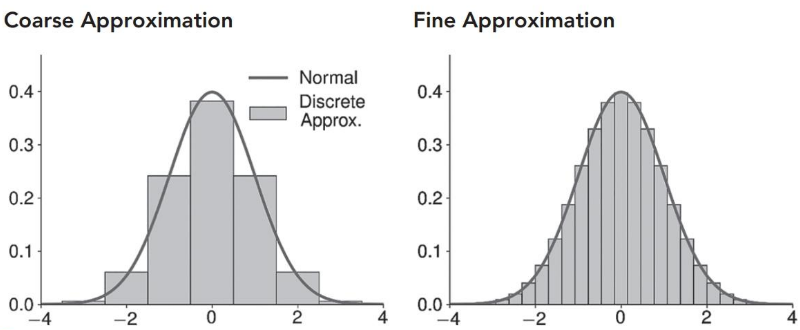


Figure 2.5 **These two graphs show discrete approximations to a continuous random variable. The left shows a coarse approximation, and the right shows a fine approximation.**

panel shows a more accurate approximation with a much smaller approximation error. As the number of points in the approximation increases, the accuracy of the approximation increases as well.

The PDF and the CDF of many common random variables are complicated. For example, the most widely used continuous distribution in finance and economics is known as a normal (or Gaussian) distribution. The PDF of a normal variable is

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{\sigma^2}\right) \quad (2.16)$$

where μ and σ^2 are parameters that determine the mean and variance of the random variable (respectively). The CDF of a normal random variable does not have a closed form expression (i.e., it lacks a formula) and is therefore calculated using numerical methods.

Fortunately, it is rare to work directly with the PDF of a continuous random variable. The next chapter describes several common continuous random variables that are used throughout finance and risk management. These common distributions have well-documented properties, and so it is not necessary to directly compute moments or most other required measures.

2.6 QUANTILES AND MODES

Moments are the most common measures used to describe and compare random variables. They are well understood and capture important features. However, quantiles can be used to construct an alternative set of descriptive measures of a random variable.

Quantiles are the cutoffs that separate the observations of a variable into (almost) equally sized buckets. For a continuous or discrete random variable X, the α-quantile X is the smallest number q such that $\Pr(X < q) = \alpha$. This is traditionally denoted by:

$$\inf_q \{q: \Pr(X \leq q) = \alpha\}, \quad (2.17)$$

where $\alpha \in [0,1]$ and inf selects the smallest possible value of q satisfying the requirement that $\Pr(X \leq q) = \alpha$. The inf is only needed when measuring the quantile of distributions that have regions with no probability.

In most applications in finance and risk management, the assumed distributions are continuous and without regions of zero probability. This means that for each α, there is exactly one number q such that $\Pr(X \leq q) = \alpha$. In this important case, the quantile function can be defined as such that $Q(\alpha)$ returns the α-quantile q of X and:

$$Q_X(\alpha) = F_X^{-1}(\alpha), \quad (2.18)$$

so that the quantile function is the inverse of the CDF. Recall that the CDF transforms values in the support of X to the cumulative probability. These values are always between 0 and 1. The quantile function thus maps a cumulative probability to the corresponding quantile.

Figure 2.6 shows the relationship between the CDF and the quantile function of a continuous random variable. The left panel shows that the CDF $F_X(x)$ maps realizations to the associated quantiles. The right panel shows how the quantile function $Q_X(\alpha) = F_X^{-1}(\alpha)$

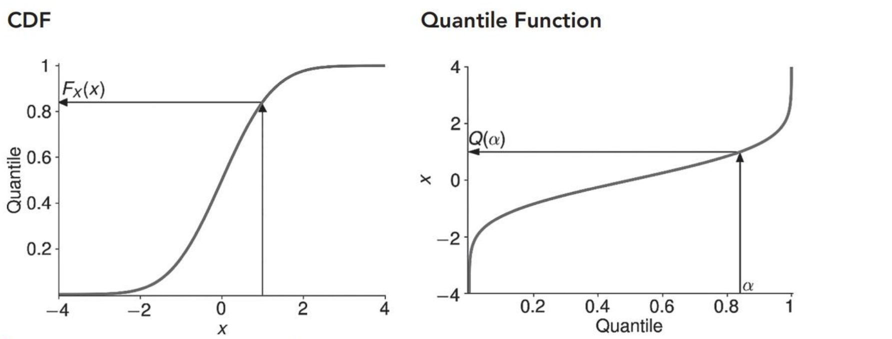


Figure 2.6 **The left panel shows the CDF of a continuous random variable, which maps values x to the associated quantile (i.e., $F_X(x)$). The right panel shows the quantile function, which maps quantiles to the corresponding inverse CDF value (i.e., $Q_X(\alpha) = F_x^{-1}(\alpha)$). It can easily be seen that either graph is a reflection of the other through the diagonal (i.e., with the axes flipped).**

maps quantiles to the values where $\Pr(X \leq x) = \alpha$. This means that the quantile function is the reflection of the CDF along a 45° line.

Quantiles are used to construct useful descriptions of random variables that complement moments. Two are common: the median and the interquartile range.

The median is defined as the 50% quantile and measures the point that evenly splits the probability in the distribution (i.e., so that half of the probability is below the median and half is above it). The median is a measure of location and is often reported along with the mean. When distributions are symmetric, the mean and the median are identical. When distributions are skewed, these values generally differ (e.g., in the common case of negative skew, the mean is below the median). The interquartile range (IQR) is defined as the difference between the 75% and 25% quantiles. It is a measure of dispersion that is comparable to the standard deviation.

Both the IQR and the median depend on values that are relatively central and never in the tail of the distribution. This makes them less sensitive to changes in the extreme tail of a distribution than the standard deviation and mean, respectively.

Quantiles are also noteworthy because they are always well-defined for any random variable. Moments, even the mean, may not be finite when a random variable is very heavy-tailed. The median, or any other quantile, is always well-defined even in cases where the mean is not. The top panels of Figure 2.7 show the relationship between the mean and the median in a

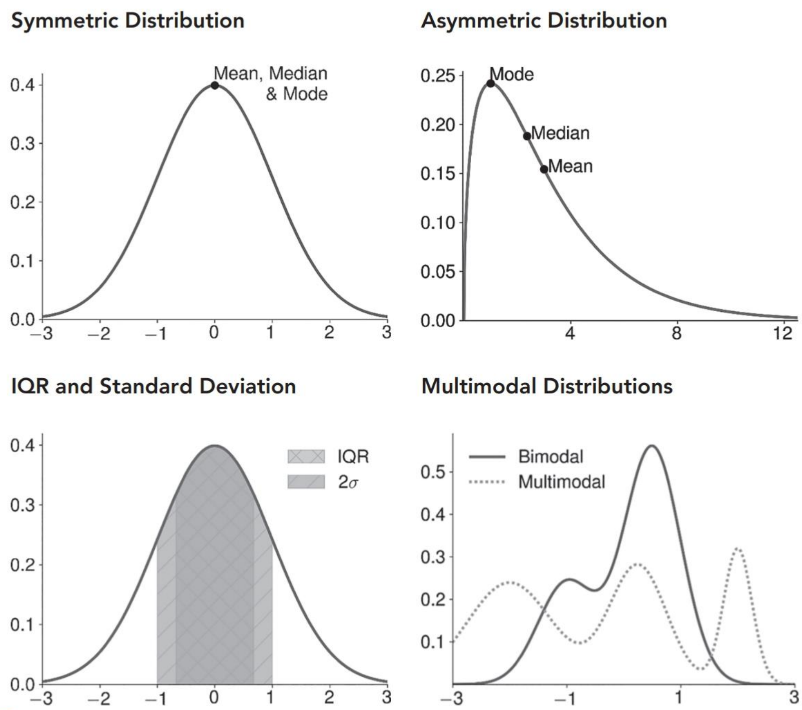


Figure 2.7 **The top panels demonstrate the relationship between the mean, median, and mode in a symmetric, unimodal distribution (left) and a unimodal right-skewed distribution (right). The bottom-left panel demonstrates the relationship between the interquartile range and the standard deviation. The bottom-right panel contains examples of both a bimodal and a multimodal distribution.**

symmetric distribution (left) and in a positively skewed distribution (right). The bottom-left panel relates the IQR to a 2σ interval in a normal distribution.

One final measure of central tendency, the mode, is also frequently used to characterize the distribution of a random variable. Specifically, the mode is a measure of *common* tendency. It measures the location of the most frequently observed values from a random variable. When a random variable is continuous, the mode is the highest point in the PDF.

Note that random variables may have one or more modes. When a distribution has more than one mode, it is common to label any local maximum in the PDF as a mode. These distributions are described as bimodal (if there are two modes) or multimodal (if there are more than two modes). The bottom-right panel of Figure 2.7 shows a bimodal and a multimodal distribution.

2.7 SUMMARY

This chapter introduced random variables, which are functions from the sample space that return numeric values. This property allows us to apply standard results from mathematics to characterize their behavior. The probability of observing various values of a random variable is summarized by the probability mass function (PMF) and the cumulative distribution function (CDF). These two functions are closely related: The PMF returns the probability of a specific value x, while the CDF returns the total probability that the random variable is less than or equal to x.

Expectations are a natural method to compute the average value of a random variable (or a function of a random variable). When a random variable X is discrete, the expectation is just a sum of the outcomes weighted by the probabilities of each outcome.

Moments are the expected values for simple functions of a random variable. The four key moments are:

1. The mean, which measures the average value of the random variable;
2. The variance, which is a measure of deviation from the mean;
3. The skewness, which measures asymmetry; and
4. The kurtosis, which measures the extent to which a variable's distribution contains extreme values.

Financial variables are often rescaled, and these four moments all change in easily understandable ways:

- The mean is affected by location shifts and rescaling.
- The variance is only affected by rescaling.
- The skewness and kurtosis are unaffected by increasing linear transformations because they are unit-free.

Finally, the quantile function allows two other common measures to be defined:

1. The median, which is an alternative measure of central tendency.
2. The interquartile range, which measures dispersion.

Later chapters show how these measures—both moments and quantiles—are used to enhance our understanding of observed data.

The following questions are intended to help candidates understand the material. They are *not* actual FRM exam questions.

QUESTIONS

Short Concept Questions

2.1 What is the key difference between a discrete and a continuous random variable?

2.2 What do a PMF and a CDF measure? How are these related?

2.3 What two properties must a PMF satisfy?

2.4 How are the mean, variance, and standard deviation of Y related to these moments of X when $Y = a + bX$? In other words, calculate the mean, variance, and standard deviation of Y as a function of the mean, variance, and standard deviation of X.

2.5 What is excess kurtosis?

2.6 How are quantile functions related to the CDF? When X is a continuous random variable, when does this relationship not hold?

2.7 What are the median and interquartile range? When is the median equal to the mean?

Practice Questions

2.8 What is $E[XE[X]]$? Hint: recall that $E[X]$ is a constant.

2.9 Consider the following data on a discrete random variable X:

	X
1	−2.456
2	−3.388
3	−6.816
4	1.531
5	1.737
6	−1.254
7	−1.164
8	1.532
9	2.550
10	0.296
11	−0.979
12	−4.259
13	2.810
14	−1.608
15	−0.575

a. Calculate the mean and variance of X.
b. Standardize X and check that the mean and variance of X are 0 and 1 respectively.
c. Calculate the skew and kurtosis of X.

2.10 Suppose the return on an asset has the following distribution:

Return	Probability
−4%	6%
−3%	9%
−2%	11%
−1%	12%
0%	14%
1%	16%
2%	15%
3%	8%
4%	5%
5%	4%

a. Compute the mean, variance, and standard deviation.
b. Verify your result in (a) by computing $E[X^2]$ directly and using the alternative expression for the variance.
c. Is this distribution skewed?
d. Does this distribution have excess kurtosis?
e. What is the median of this distribution?

The following questions are intended to help candidates understand the material. They are *not* actual FRM exam questions.

ANSWERS

Short Concept Questions

2.1 Discrete random variables have support on only a finite (or possibly countably infinite) set of points. These points are all distinct in the sense that it there is always some number between any two points in the support. A continuous random variable has support over a continuous range of values. This range can be finite, as in a uniform random variable; or infinite, as in a standard normal random variable.

2.2 The PMF measures the probability of a single value. The CDF measures the total probability that the random variable is less than or equal to a specific value—that is, the cumulative probability that $X \le x$ for some value x. The CDF is the sum of the PMF across all values less than x.

2.3 A PMF must be non-negative and sum across the support of the random variable to 1.

2.4 How are the mean, variance, and standard deviation of Y related to these moments of X when $Y = a + bX$? In other words, calculate the mean, variance, and standard deviation of Y as a function of the mean, variance and standard deviation of X

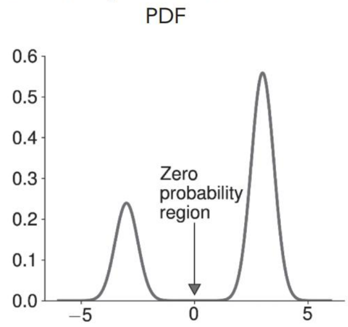
$E[Y] = E[a + bX]$
$= E[a] + E[bX]$ Expectation of sum is the sum of expectations
$= a + bE[X]$ Expectation of a constant is constant.

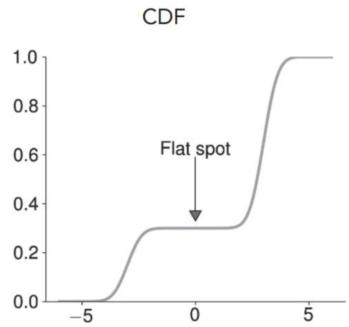
The variance of Y is $V[a + bX] = V[bX] = b^2V[X]$ because the variance of a constant is 0 and the variance of bX is b^2 times the variance of X.

The standard deviation of Y is simply the square root of the variance given above, $\sqrt{b^2V[X]} = |b|\sqrt{V[X]}$.

2.5 Excess kurtosis is the standard kurtosis measure minus 3. This reference value comes from the normal distribution, and so excess kurtosis measures the kurtosis above that of the normal distribution.

2.6 The quantile function is the inverse of the CDF function. That is, if $u = F_X(x)$ returns the CDF value of X, then $q = F_X^{-1}(u)$ is the quantile function. The CDF function is not invertible if there are regions where there is no probability. This corresponds to a *jump* in the CDF, as shown in the figures that follow.





2.7 The median is the point where 50% of the probability is located on either side. It may not be unique in discrete distributions, although it is common to choose the smallest value where this condition is satisfied. The IQR is the difference between the 25% and 75% quantiles. These are the points where 25% and 75% of the probability of the random variable lies to the left. This difference is a measure of spread that is like the standard deviation.

The median is equal to the mean in any symmetric distribution if the first moment exists (is finite).

The following questions are intended to help candidates understand the material. They are *not* actual FRM exam questions.

Solved Problems

2.8 $E[XE[X]] = \Sigma E[X]X \Pr(X = x) = E[X]\Sigma X \Pr(X = x) = E[X] \times E[X] = (E[X])^2$ because $E[X]$ is a constant and does not depend on X. In other words, the expected value of a random variable *does not* depend on the random variable.

2.9 **a.** The mean is −0.803 and the variance is 6.762.

b.

	X	X Standardized
1	−2.456	−0.636
2	−3.388	−0.994
3	−6.816	−2.312
4	1.531	0.897
5	1.737	0.977
6	−1.254	−0.173
7	−1.164	−0.139
8	1.532	0.898
9	2.550	1.289
10	0.296	0.423
11	−0.979	−0.068
12	−4.259	−1.329
13	2.810	1.389
14	−1.608	−0.310
15	−0.575	0.088
Mean	−0.803	0.000
Variance	6.762	1.000
Standard Deviation	2.600	1.000

c.

	X	X Standardized	(X Standardized)^3 (Skew)	(X Standardized)^4 (Kurtosis)
1	−2.456	−0.636	−0.257	0.163
2	−3.388	−0.994	−0.982	0.977
3	−6.816	−2.312	−12.364	28.590
4	1.531	0.897	0.723	0.649
5	1.737	0.977	0.932	0.910
6	−1.254	−0.173	−0.005	0.001
7	−1.164	−0.139	−0.003	0.000
8	1.532	0.898	0.724	0.650
9	2.550	1.289	2.143	2.764
10	0.296	0.423	0.075	0.032
11	−0.979	−0.068	0.000	0.000
12	−4.259	−1.329	−2.348	3.120
13	2.810	1.389	2.682	3.726
14	−1.608	−0.310	−0.030	0.009
15	−0.575	0.088	0.001	0.000
Mean	−0.803	0.000	−0.581	2.773
Variance	6.762	1.000		
Standard Deviation	2.600	1.000		

Hence, the skew of X is −0.581 and the kurtosis of X is 2.773.

The following questions are intended to help candidates understand the material. They are *not* actual FRM exam questions.

2.10 **a.** The mean is:

$$E[X] = \Sigma x \Pr(X = x)$$
$$= -4\% \times 6\% + -3\% \times 9\% + \ldots + 5\% \times 4\%$$
$$= 0.25\%.$$

The variance is:

$$Var[X] = \Sigma(x - E[X])^2 \Pr(X = x)$$
$$= (-4\% - 0.25\%)^2 \times 6\% + (-3\% - 0.25\%)^2 \times 9\% + \ldots + (5\% - 0.25\%)^2 \times 4\% = 0.0555\%.$$

The standard deviation is $\sqrt{Var[X]} = 2.355\%$.

b. $E[X^2] = \Sigma x^2 \Pr(X = x) = .000561$ and so $E[X^2] - (E[X])^2 = 0.000561 - (.0025)^2 = .000555$, which is the same.

c. The skewness requires computing

$$\text{skew}(X) = \frac{E[(X - E[X])^3}{\sigma^3} = E\left[\left(\frac{X - \mu}{\sigma}\right)^3\right]$$
$$= \sum\left(\frac{x - \mu}{\sigma}\right)^3 \Pr(X = x).$$

Thus the skewness is 0.021, and the distribution has a mild positive skew.

d. The kurtosis requires computing

$$\text{kurtosis}(X) = \frac{E[(X - E[X])^4}{\sigma^4} = E\left[\left(\frac{X - \mu}{\sigma}\right)^4\right]$$
$$= \sum\left(\frac{x - \mu}{\sigma}\right)^4 \Pr(X = x).$$

Thus the kurtosis is 2.24. The excess kurtosis is then $2.24 - 3 = -0.76$. This distribution does not have excess kurtosis.

e. The median is the value where at least 50% probability lies to the left, and at least 50% probability lies to the right. Cumulating the probabilities into a CDF, this occurs at the return value of 0%:

Return	Probability	Cumulative Probability
−4%	6%	6%
−3%	9%	15%
−2%	11%	26%
−1%	12%	38%
0%	14%	52%
1%	16%	68%
2%	15%	83%
3%	8%	91%
4%	5%	96%
5%	4%	100%

3 Common Univariate Random Variables

Learning Objectives

After completing this reading you should be able to:

- Distinguish the key properties and identify the common occurrences of the following distributions: uniform distribution, Bernoulli distribution, binomial distribution, Poisson distribution, normal distribution, lognormal distribution, Chi-squared distribution, Student's *t* and *F*-distributions.
- Describe a mixture distribution and explain the creation and characteristics of mixture distributions.

There are over two hundred named random variable distributions. Each of these distributions has been developed to explain key features of real-world phenomena. This chapter examines a set of distributions commonly applied to financial data and used by risk managers.

Risk managers model uncertainty in many forms, so this set includes both discrete and continuous random variables. There are three common discrete distributions: the Bernoulli (which was introduced in the previous chapter), the binomial, and the Poisson. The Bernoulli is a general purpose distribution that is typically used to model binary events. The binomial distribution describes the sum of n independent Bernoulli random variables. The Poisson distribution is commonly used to model *hazard rates*, which count the number of events that occur in a fixed unit of time (e.g., the number of corporations defaulting in the next quarter).

There is a wider variety of continuous distributions used by risk managers. The most basic is a uniform distribution, which serves as a foundation for all random variables. The most widely used distribution is the normal, which is used for tasks such as modeling financial returns and implementing statistical tests. Many other frequently used distributions are closely related to the normal. These include the Student's t, the chi-square (χ^2), and the F, all of which are encountered when evaluating statistical models.

This chapter concludes by introducing mixture distributions, which are built using two or more distinct component distributions. A mixture is produced by randomly sampling from each component so that the mixture distribution inherits characteristics of each component. Mixtures can be used to build distributions that match important features of financial data. For example, mixing two normal random variables with different variances produces a random variable that has a larger kurtosis than either of the mixture components.

3.1 DISCRETE RANDOM VARIABLES

Bernoulli

The Bernoulli is a simple and frequently encountered distribution. It is a discrete distribution for random variables that produces one of two values: 0 or 1. It applies to any problem with a binary outcome (e.g., bull and bear markets, corporate defaults, or the classification of fraudulent transactions). It is common to label the outcome 1 as a 'success' and 0 as a 'failure'. These labels are often misnomers in finance and risk management because 1 is used to denote the *unusual* state, which might be undesirable (e.g., a bear market, a severe loss in a portfolio, or a company defaulting on its obligations).

The Bernoulli distribution depends on a single parameter, p, which is the probability that a success (i.e., 1) is observed. The probability mass function (PMF) of a Bernoulli(p) is

$$f_Y(y) = p^y(1-p)^{(1-y)} \tag{3.1}$$

Note that this function only produces two values, which are obtained by plugging values of y into the function:

$$p, \text{ when } y = 1$$

$$1 - p, \text{ when } y = 0$$

Meanwhile, its CDF is a step function with three values:

$$F_Y(y) = \begin{cases} 0 & y < 0 \\ 1-p & 0 \le y < 1 \\ 1 & y \ge 1 \end{cases} \tag{3.2}$$

The moments of a Bernoulli random variable are easy to calculate. Suppose that Y is a Bernoulli random variable with parameter p. This can be expressed as

$$Y \sim \text{Bernoulli}(p)$$

where ~ means "is distributed as." The mean and variance of Y are respectively:

$$E[Y] = p \times 1 + (1-p) \times 0 = p,$$

and

$$\begin{aligned} V[Y] = E[Y^2] - E[Y]^2 &= (p \times 1^2 + (1-p) \times 0^2) - p^2 \\ &= p - p^2 = p(1-p) \end{aligned}$$

Note that the mean reflects the probability of observing a 1, and thus it is just p. The variance reflects the fact that if p is either near 0 or near 1, then one value is observed much more frequently than the other (and thus the variance is relatively low). The variance is maximized at $V[Y] = 25\%$ when $p = 50\%$, which reflects the equal probability that failure and success (i.e., 0 and 1) are observed.

Figure 3.1 describes two Bernoulli random variables: $Y \sim \text{Bernoulli}(p = 0.5)$ and $Y \sim \text{Bernoulli}(p = 0.9)$. Their PMFs and CDFs are shown in the left and right panels, respectively. Note that the CDFs are step functions that increase from 0 to $1 - p$ at 0, and then to 1 at 1.

Bernoulli random variables can be applied to events that are binary in nature. For example, loan defaults can be modeled by making the observed values 1 for a default and 0 for no default. Meanwhile, p (i.e., the probability of default) can be calculated using various relevant factors. Complete models can include factors specific to each borrower (e.g., a credit score) as well as those that reflect wider economic conditions (e.g., the GDP growth rate).

Bernoulli random variables can also be used to model exceptionally large portfolio losses, taking on a value of 1 when such a loss is realized (e.g., a loss that exceeds Value-at-Risk [VaR]) and 0 otherwise.

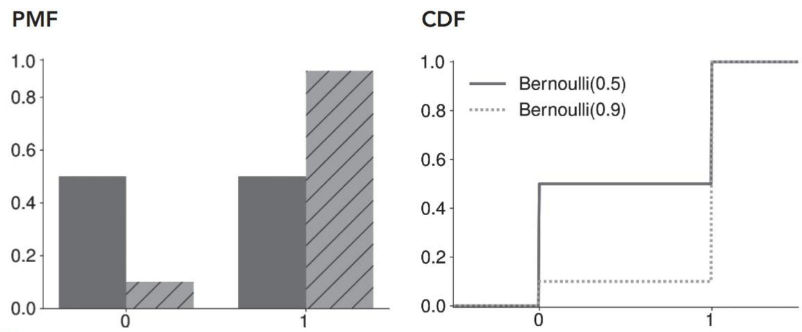


Figure 3.1 The left panel shows the probability mass functions of two Bernoulli random variables, one with p = 0.5 and the other with p = 0.9. The right panel shows the cumulative distribution functions for the same two variables.

Binomial

A binomial random variable measures the total number of successes from n independent Bernoulli random variables, where each has a probability of success equal to p. In other words, binomial distributions are used to model counts of independent events.

A binomial distribution has two parameters:

1. n, the number of independent experiments; and
2. p, the probability that each experiment is successful.[1]

If n variables $X_i \sim$ Bernoulli(p) are independent, then a binomial with parameters n and p is defined as $Y = \sum_{i=1}^{n} X_i$ and expressed as $B(n, p)$.

The mean follows directly from the properties of moments described in the previous chapter

$$E[Y] = \sum_{i=1}^{n} = p = p + p + \ldots + p = np$$

Meanwhile, the variance of Y is:

$$V[Y] = np(1 - p)$$

Note that the mean and variance for a Binomial are simply the mean and variance for a Bernoulli multiplied by n, since each of the n Bernoulli variances is independent. The $p(1 - p)$ component reflects the uncertainty coming from the Bernoulli random variables, while n considers the number of independent component variables.

[1] The outcomes of Bernoulli random variables are called experiments due to the historical application in a laboratory setting. While financial data are not usually experimental, this term is still used.

By construction, a binomial random variable is always non-negative, integer-valued, and a $B(n, p)$ is always less than or equal to n. The skewness of a binomial depends on p, with small values producing right-skewed distributions. The PMF of a $B(n, p)$ is

$$f_Y(y) = \binom{n}{y} p^y (1 - p)^{n-y} \tag{3.3}$$

Note that $\binom{n}{y}$, which is commonly expressed as "n choose y," counts the number of distinct ways that y successes could have been realized from n experiments. It is equivalent to:

$$\binom{n}{y} = \frac{n!}{y!(n - y)!},$$

where "!" indicates factorial.[2]

The probability mass function (PMF) for any values of y and n can be computed using basic probability. For example, consider a binomial random variable that models the flipping of two coins. In this case, label an outcome of heads to represent a success (i.e., it takes on a value of 1), whereas a tails outcome will be considered a failure (i.e., it takes on a value of 0). This variable can be expressed as $Y \sim B(2, 0.5)$. Note that:

- $n = 2$ because there are two flips, and
- $p = 0.5$ because each flip has a 50% chance of success (i.e., producing heads).

When the two coins are flipped, there are four possible outcomes: {Tails, Tails}, {Tails, Heads}, {Heads, Tails} or {Heads, Heads}. This means that the number of successes (i.e., heads) can either be 0, 1, or 2.

[2] $n! = n \times n - 1 \times n - 2 \times \ldots \times 2 \times 1$ for n a positive integer ≥ 1 (e.g., $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$).

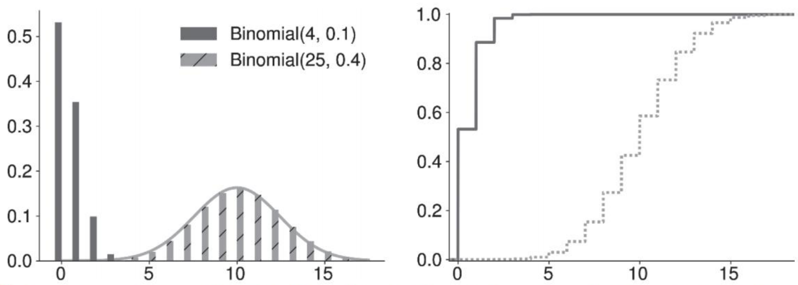


Figure 3.2 **The left panel shows the PMF of two binomial distributions. The solid line is the PDF of a normal random variable, N(np, np(1−p)) = N(10, 6), that approximates the PMF of the B(25, 0.4). The right panel shows the CDFs of the two binomials.**

Since the coin flips are independent, the probability of {Tails, Tails} is the product of the probability of each flip being tails, which is:

$$(1/2)^2 = 1/4$$

Furthermore, the probability of one head being followed by one tail is also 1/4, as is the probability of two heads.

Now, the variable's PDF can be produced by multiplying the probabilities of each number of by the number of ways in which it could occur

$$F_Y(y) = \begin{cases} 1 \times {}^1\!/_4 = \binom{2}{0} \times ({}^1\!/_2)^0 \times ({}^1\!/_2)^2 = 1/4 & \text{if } y = 0 \\ 2 \times {}^1\!/_4 = \binom{2}{1} \times ({}^1\!/_2)^1 \times ({}^1\!/_2)^1 = 1/2 & \text{if } y = 1. \\ 1 \times {}^1\!/_4 = \binom{2}{2} \times ({}^1\!/_2)^2 \times ({}^1\!/_2)^0 = 1/4 & \text{if } y = 2 \end{cases}$$

Meanwhile, the CDF is the sum[3] of the cumulated PMF between 0 and y

$$F_Y(y) = \sum_{i=0}^{\lfloor y \rfloor} \binom{n}{i} p^i (1-p)^{n-i} \tag{3.4}$$

The CDF is defined using the floor function $\lfloor y \rfloor$, which returns y when y is an integer and the largest integer smaller than y when y is not.

Note that the normal distribution provides a convenient approximation to the binomial if both $np \geq 10$ and $n(1-p) \geq 10$. This is illustrated in Figure 3.2 and explained in more detail later in this chapter.

Figure 3.2 shows the PMFs and CDFs of two binomial random variables. One variable has $n = 4$ and $p = 0.1$, which produces a distribution with right-skew. Note that the mass is heavily concentrated around $y = 0$ and $y = 1$, which reflects the small number (four) of Bernoulli random variables. Meanwhile, the second configuration has $n = 25$ and $p = 0.4$, creating a shape that resembles the density of a normal random variable. The solid line in the PMF panel shows the PDF of this approximation. The right panel shows the CDFs, which are step functions (since binomials are discrete random variables).

Poisson

Poisson random variables are used to measure counts of events over fixed time spans. For example, one application of a Poisson is to model the number of loan defaults that occur each month. Poisson random variables are always non-negative and integer-valued. The Poisson distribution has a single parameter, which is called the hazard rate and expressed as λ, that signifies the average number of events per interval. Therefore, the mean and variance of $Y \sim \text{Poisson}(\lambda)$ is simply:

$$E[Y] = V[Y] = \lambda$$

The PMF of a Poisson random variable is

$$f_Y(y) = \frac{\lambda^y \exp(-\lambda)}{y!} \tag{3.5}$$

Meanwhile, the CDF of a Poisson is defined as the sum of the PMF for values less than the input

$$F_Y(y) = \exp(-\lambda) \sum_{i=0}^{\lfloor y \rfloor} \frac{\lambda^i}{i!} \tag{3.6}$$

The Poisson parameter λ can be considered as a hazard rate in survival modeling.

3

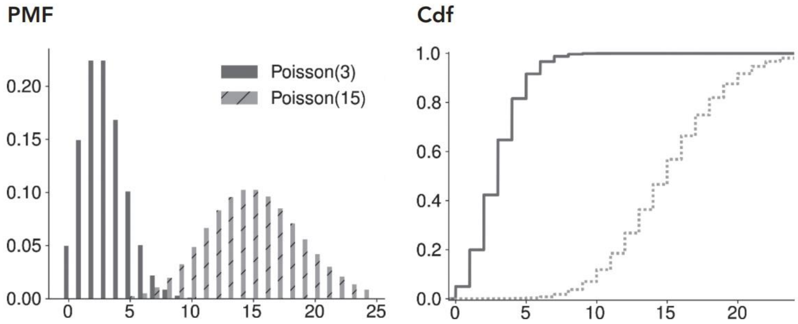


Figure 3.3 **The left panel shows the PMF for two values of λ, 3 and 15. The right panel shows the corresponding CDFs.**

As an example, consider a fixed-income portfolio that consists of a large number of bonds. On average, five bonds within the portfolio default each month. Assuming that the probability of any bond defaulting is independent of the other bonds, what is the probability that exactly two bonds default in one month?

In this case, $y = 2$ and $\lambda = 5$ since the mean number of defaults within a month is five and y is the value being tested. Thus

$$f_Y(2) = \frac{5^2 \exp(-5)}{2!} = 0.0842 = 8.42\%$$

A useful feature of the Poisson (one that is uncommon among discrete distributions) is that it is infinitely divisible. If $X_1 \sim \text{Poisson}(\lambda_1)$, and $X_2 \sim \text{Poisson}(\lambda_2)$ are independent, and $Y = X_1 + X_2$, then $Y \sim \text{Poisson}(\lambda_1 + \lambda_2)$; in other words, the sum of independent Poisson variables is also a Poisson. This feature makes the distribution well-suited to work with time-series data, because summing the number of events in a sampling interval (e.g., a week, month, or quarter) does not change the distribution.

Figure 3.3 shows the PMFs (left panel) and CDFs (right panel) for two configurations of a Poisson random variable. When λ is small, the distribution is right-skewed. When $\lambda = 15$, the distribution is nearly symmetric.

3.2 CONTINUOUS RANDOM VARIABLES

Uniform

The simplest continuous random variable is a uniform random variable. A uniform distribution assumes that any value within the range $[a, b]$ is equally likely to occur. The PDF of a uniform is:

$$f_Y(y) = \frac{1}{b - a} \tag{3.7}$$

The PDF of a uniform random variable does *not* depend on y, because all values are equally likely.

The CDF returns the cumulative probability of observing a value less than or equal to the argument, and is:

$$F_Y(y) = \begin{cases} 0 & \text{if } y < a \\ \dfrac{y - a}{b - a} & \text{if } a \le y < b. \\ 1 & \text{if } y \ge b \end{cases} \tag{3.8}$$

The CDF is 0 to the left of a, which is the smallest value that could be produced, linearly increases from a to b, and then is 1 above b. When $a = 0$ and $b = 1$, the distribution is called the standard uniform. Any uniform random variable can be constructed from a standard uniform U_1 using the relationship:

$$U_2 = a + (b - a)U_1, \tag{3.9}$$

where a and b are the bounds of U_2.

Figure 3.4 plots the PDFs of $U(0, 1)$ and $U(1/2, 3)$. These PDFs both have a simple box-shape, which reflects the constant probability across the support of the random variable. The right panel shows the CDFs that are piecewise linear functions that are 0 below the lower bound, 1 above the upper bound, and have a slope of $1/(b - a)$ between these two values.

The mean of a uniform random variable $Y \sim U(a, b)$ is the midpoint of the support:[4]

$$E[Y] = \frac{a + b}{2}$$

[4] Deriving the formulas for the mean and variance of a uniform distribution would require integration of y multiplied by the PDF.

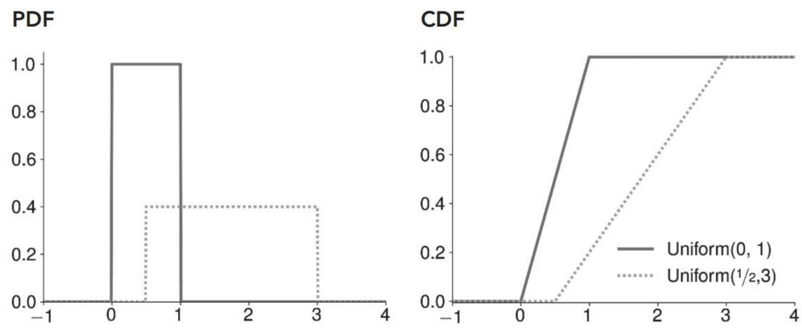


Figure 3.4 The left panel shows the probability density functions of a standard uniform (i.e., U(0, 1)) and a uniform between 1/2 and 3. The right panel shows the cumulative distribution functions corresponding to these two random variables.

Meanwhile, the variance depends only on the width of the support:

$$V[Y] = \frac{(b-a)^2}{12}$$

When $a = 0$ and $b = 1$, the mean and variance of a standard uniform random variable are 1/2 and 1/12, respectively.

The probability that a uniform(a, b) random variable Y falls into an interval with lower bound l and upper bound u (i.e., $\Pr(l < Y < u)$) is

$$\frac{\min(u, b) - \max(l, a)}{b - a}, \quad (3.10)$$

which simplifies to $(u - l)/(b - a)$ when both $l \geq a$ and $u \leq b$.

As an example, a uniform random variable $Y \sim U(-2, 4)$ has a mean of

$$E[Y] = \frac{-2+4}{2} = 1$$

A variance of

$$V[Y] = \frac{(4-(-2))^2}{12} = 3$$

And the probability that Y falls between −1 and 2 is

$$\Pr(l < Y < u) = \frac{\min(2, 4) - \max(-1, -2)}{4-(-2)} = \frac{2-(-1)}{4-(-2)} = 0.5$$

Normal

The normal distribution is the most commonly used distribution in risk management. It is also commonly referred to as a Gaussian distribution (after mathematician Carl Friedrich Gauss) or a bell curve (which reflects the shape of the PDF). The normal distribution is popular for many reasons.

- Many continuous random variables are approximately normally distributed.
- The distribution of many discrete random variables can be well approximated by a normal.
- The normal distribution plays a key role in the Central Limit Theorem (CLT), which is widely used in hypothesis testing (i.e., the process where data are used to determine the truthfulness of an objective statistical statement).
- Normal random variables are *infinitely divisible*, which makes them suitable for simulating asset prices in models that assume that prices are continuously evolving.
- The normal is closely related to many other important distributions, including the Student's t, the χ^2, and the F.
- The normal is closed (i.e., weighted sums of normal random variables are normally distributed) under linear operations.
- Estimators derived under the assumption that the underlying observations are normally distributed often have simple closed forms.

The normal distribution has two parameters: μ (i.e., the mean) and σ^2 (the variance). Therefore

$$E[Y] = \mu$$

and

$$V[Y] = \sigma^2$$

A normal distribution has no skewness (because it is symmetrical) and a kurtosis of 3. This kurtosis is often used as a benchmark when assessing whether another distribution is *heavy/fat-tailed* (i.e., has relatively greater probability of observing an especially large deviation than if the random variable is normally distributed).

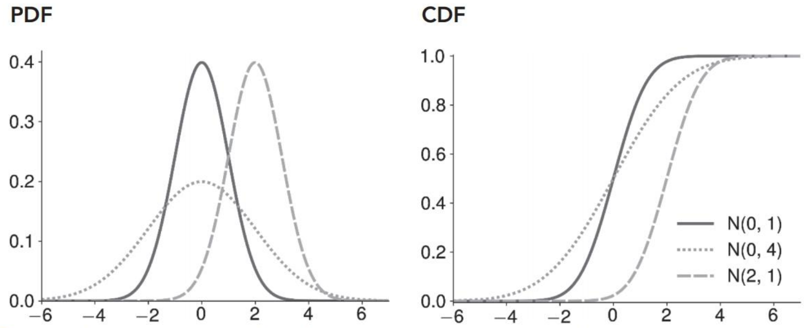


Figure 3.5 **The left panel shows the PDFs of three normal random variables, N(0, 1) (i.e., the standard normal), N(0, 4), and N(2, 1). The right panel shows the corresponding CDFs.**

Figure 3.5 shows the shapes of the PDFs of three normal distributions, all of which have a bell shape. Each PDF is centered at the respective mean μ of each distribution. Note that increasing the variance σ^2 increases the spread of a distribution around its mean.

The normal can generate any value in $(-\infty, \infty)$, although it is unlikely (i.e., less than one in 10,000) to observe values more than 4σ away from the mean. In fact, values more than 3σ away from the mean are expected in only one in 370 realizations of a normal random variable.

The PDF of a normal is

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \quad (3.11)$$

Recall that while the CDF of a normal does not have a closed form, fast numerical approximations are widely available in Excel or other statistical software packages.

An important special case of a normal distribution occurs when $\mu = 0$ and $\sigma^2 = 1$. This is called a standard normal, and it is common to use Z to denote a random variable with distribution $N(0,1)$. It is also common to use $\phi(z)$ to denote the standard normal PDF and $\Phi(z)$ to denote the standard normal CDF.

The normal is widely applied, and key quantiles from the normal distribution are commonly used for two purposes.

1. The quantiles are used to approximate the chance of observing values more than 1σ, 2σ, and 3σ when describing a log return as a normal random variable.
2. When constructing confidence intervals, which provide a range of values within which the true but unknown parameter value is likely to be, it is common to consider symmetric intervals constructed using a standard normal distribution that contain 90%, 95%, or 99% of the probability.

These values for a normal distribution are summarized in Table 3.1. The left panel tabulates the area covered by an interval of $q\sigma$ for $q = 1, 2, 3$. Both the exact probabilities and the common approximations used by practitioners are given. For each example, the table shows that exactly 95.4% of the probability mass of a normal distribution (the area under the PDF) lies within two standard deviations of the mean while 99.7% of it lies within three standard deviations.

The right panel shows the end points of a symmetric interval with a total tail probability of α. This means that each tail contains probability $\alpha/2$ for $\alpha = 10\%$, 5%, and 1%. The central probability of each region is $100\% - \alpha$. It is common to substitute 2 for 1.96 here as well. In other words, 2.5% of the probability mass of a standard normal lies to the right of 1.96 and 2.5% lies to the left of −1.96, with 95% of it between −1.96 and 1.96.

Figure 3.6 shows a graphical representation of these two measures. The left panel shows the area covered by 1σ, 2σ, and 3σ. These regions are centered at the mean μ. The right panel shows the area covered by three confidence intervals when the random variable is a standard normal.

Note that the sums of independent[5] normally distributed random variables are also normally distributed. For instance, if $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ are independent, and $Y = X_1 + X_2$, then $Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. This property simplifies describing log returns at different frequencies; if daily log returns are independent and normal, then weekly and monthly returns are as well.

[5] Chapter 4 shows that property also holds more generally for normal random variables and independence is not required.

Table 3.1 **The Left Panel Shows the Total Probability within 1, 2, and 3 Standard Deviations of the Mean for a Normal Random Variable. The Right Most Column of the Left Panel Contains the Common Approximation Used for Each Multiple. The Right Panel Shows the Endpoints for a Symmetric Interval with a Total Tail Probability of 10%, 5%, or 1% for a Standard Normal**

Standardized Distance from μ			Confidence Intervals		
	Exact	Common Approximation	Central Probability	Tail Probability	Value
1σ	68.2%	68%	90%	10%	± 1.645
2σ	95.4%	95%	95%	5%	± 1.96
3σ	99.7%	99%	99%	1%	± 2.57

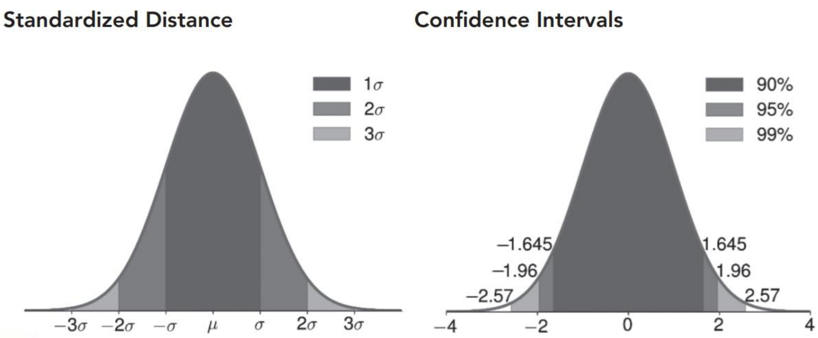


Figure 3.6 **The left panel shows the area (probability) covered by ± 1, ± 2, and $\pm 3\sigma$. The right panel shows the critical values that define regions with 90%, 95%, and 99% coverage so that the total area in the tails is 10%, 5%, and 1%, respectively. The probability in each tail is $\alpha/2$.**

Approximating Discrete Random Variables

Recall that a normal distribution can approximate a binomial random variable if both np and $n(1 - p)$ are sufficiently large. It is common to assume that the approximation will be sufficiently accurate when np and $n(1 - p)$ are greater than 10. When these conditions are satisfied, a binomial has either many independent experiments or a probability that is not extreme (or both), and so the PMF is nearly symmetric and well approximated by a $N(np, np(1 - p))$.[6]

The Poisson can also be approximated by a normal random variable. When λ is large, then a Poisson(λ) can be well approximated by a Normal(λ, λ). This approximation is commonly applied when $\lambda \geq 1000$.

Log-Normal

A variable Y is said to be log-normally distributed if the natural logarithm of Y is normally distributed. In other words, if $X = \ln Y$, then Y is log-normally distributed if and only if X is normally distributed. Alternatively, a log-normal can be defined

$$Y = \exp(X), \tag{3.12}$$

where $X \sim N(\mu, \sigma)$. An important property of the log-normal distribution is that Y can never be negative, whereas X can be negative because it is normally distributed. This log-normal property can be desirable when constructing certain models. For example, if stock prices are assumed to be normally distributed, there is a positive (although perhaps tiny) probability that the stock price becomes negative, which would not be realistic.

[6] The approximation of a binomial with a normal random variable is an application of the central limit theorem (CLT), which establishes conditions where averages of independent and identically distributed random variables behave like normal random variables when the number of samples is large. Chapter 5 provides a thorough treatment of CLTs.

WORKING WITH NORMAL RANDOM VARIABLES

Note that a normal random variable with any mean and variance can be constructed as a simple linear transformation of another normal random variable.

To see how this works, begin with the standard normal (i.e., $Z \sim N(0,1)$). Defining $X = \mu + \sigma Z$ (where μ and σ are constants) therefore produces a random variable with a $N(\mu, \sigma^2)$ distribution.

The mean and variance of X can be verified using the results from Chapter 2 to show that:

$$E[X] = \mu + \sigma E[Z] = \mu$$

and

$$Var[X] = \sigma^2 Var[Z] = \sigma^2$$

Chapter 2 also explains how a random variable is standardized, so that if $X \sim N(\mu, \sigma^2)$, then:

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

This result allows the CDF of any normally distributed random variable to be evaluated using the CDF of a standard normal. For example, suppose that $X \sim N(3,4)$, and we wanted to evaluate the CDF of X at 2, which is $\Pr(X < 2)$. Applying the standardization transformation to both X and 2 gives:

$$\Pr(X < 2) = \Pr\left(\frac{X - \mu}{\sigma} < \frac{2 - \mu}{\sigma}\right) = \Pr\left(Z < \frac{2 - 3}{2}\right)$$
$$= \Pr\left(Z < -\frac{1}{2}\right)$$

The result allows the CDF of the original normal random variable X to be computed from the CDF of the standard normal:

$$F_X(x) = \Phi\left(\frac{x - \mu}{\sigma}\right) = \int_{-\infty}^{\frac{x-\mu}{\sigma}} \phi(z)$$

Figure 3.7 provides an illustration of a PDF, where the shaded area under the curve is the CDF.

Note that only one tail is required because the standard normal distribution is symmetric around 0 so that $\Pr(Z < 0) = \Pr(Z > 0) = 0.5$. This symmetry ensures that:

$$\Pr(Z < z) = \Pr(Z > -z) = 1 - \Pr(Z < -z)$$

when $z < 0$.

For example, $\Pr(-1 < Z < 2)$ is simply:

$$\Pr(Z < 2) - \Pr(Z < -1)$$

However, because the standard Z table only includes positive values of Z, the probability that $Z < -1$ is the same as the probability that $Z > 1$, which is $1 - \Pr(Z < 1)$. Combining these two, the total probability in this region is $\Pr(Z < 2) - (1 - \Pr(Z < 1))$. Looking up these two left tail values, $\Pr(-1 < Z < 2)$ is $0.9772 - (1 - 0.8413) = 0.8185$. This method can be applied to any normal random variable X by standardizing it, so that:

$$\Pr(l < X < u) = \Pr\left(\frac{l - \mu}{\sigma} < \frac{X - \mu}{\sigma} < \frac{u - \mu}{\sigma}\right)$$
$$= \Pr\left(\frac{l - \mu}{\sigma} < Z < \frac{u - \mu}{\sigma}\right)$$

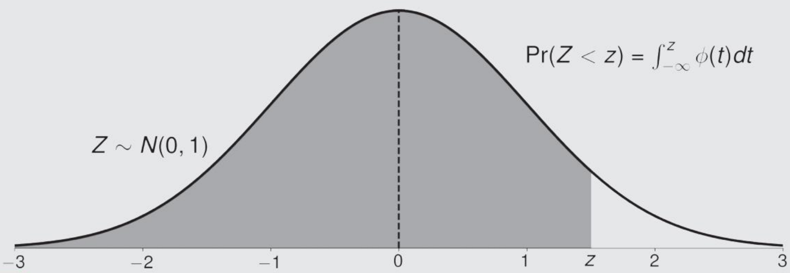


Figure 3.7 **Illustration of the PDF values of the standard normal.**

A log-normal distribution can be denoted as $Y \sim \text{Log}N(\mu, \sigma^2)$, or equivalently as $\ln(Y) \sim N(\mu, \sigma^2)$, where $\ln(Y)$ is normally distributed with mean μ and variance σ^2. Here, Y is a non-linear transformation of a normal random variable and the mean[7] of Y is $\exp(\mu + \sigma^2/2)$. This is impossible under a log-normal model and so is more appropriate than a normal for stock prices.

[7] The mean reflects the mean of the underlying normal random variable as well as its variance. The extra term arises due to Jensen's inequality: The exponential function is convex and so it must be the case that $E[Y] \geq \exp(E[X]) = \exp(\mu)$. The variance depends on both model parameters and $V[Y] = (\exp(\sigma^2) - 1)\exp(2\mu + \sigma^2)$.

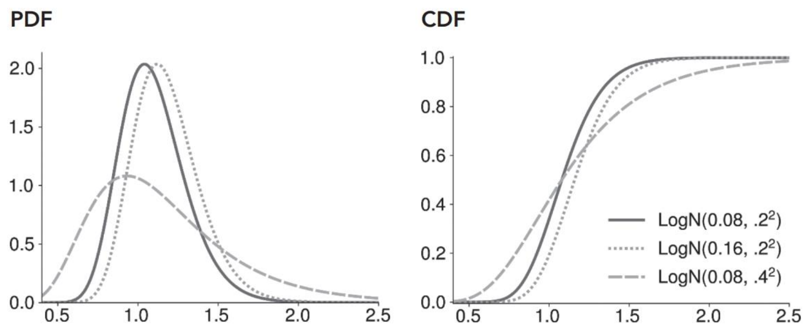


Figure 3.8 **The left panel shows the PDFs of three log-normal random variables, two of which have $\mu = 8\%$, and two with $\sigma = 20\%$, which are typical of annual equity returns. The right panel shows the corresponding CDFs.**

The PDF of a log-normal is given by:

$$f_Y(y) = \frac{1}{y\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\ln y - \mu)^2}{2\sigma^2}\right) \tag{3.13}$$

Because a log-normal random variable is defined as the exponential of a normal random variable [i.e., $Y = \exp X$, where $X \sim N(\mu, \sigma^2)$], the CDF of Y can be computed by inverting the transformation using $\ln(Y)$ and then applying the normal CDF on the transformed value

$$F_Y(y) = \Phi\left(\frac{\ln y - \mu}{\sigma}\right) \tag{3.14}$$

Figure 3.8 shows the PDFs (left panel) and CDFs (right panel) for three log-normal random variables. These have been parameterized to reflect plausible distributions of annual returns on an equity market index. Note that the expected value of the distribution is always greater than $\exp(\mu)$ due to the $\sigma^2/2$ term in the mean. For example, when $\mu = 8\%$ and $\sigma = 20\%$, the expected value is 1.105, and when $\sigma = 40\%$, the expected value is 1.174. At the same time, changing μ to 16% produces an expected value of 1.197. The effect of σ^2 on the expected value reflects the asymmetric nature of returns; the worst return is −100%, whereas the upside is unbounded. Since the exponential transform is nonlinear, the log-normal distribution Y is not symmetric even though the normal distribution X is.

χ^2

The χ^2 (chi-squared) distribution is frequently encountered when testing hypotheses about model parameters. It is also used when modeling variables that are always positive, (e.g., the VIX Index). A χ^2_ν random variable can be defined as the sum of the squares of ν (i.e., the Greek letter nu) independent standard normal random variables:

$$Y = \sum_{i=1}^{\nu} Z_i^2 \tag{3.15}$$

A χ^2 distribution is considered to have ν *degrees of freedom*, which in this case is a positive integer parameter ($\nu = 1, 2, \ldots$) that defines the shape of the distribution. Degrees of freedom measure the amount of data that is available to test model parameters, because estimating model parameters requires a minimum number of observations (e.g., k). In many models, the degree of freedom used in testing is $n - k$.

Using standard properties for sums of independent random variables, it can be shown that for $Y \sim \chi^2_\nu$

$$E[Y] = \nu$$

and

$$V[Y] = 2\nu$$

The PDF of a χ^2_ν random variable is:

$$f_Y(y) = \frac{1}{2^{(\nu/2)}\Gamma(\nu/2)} y^{(\nu/2 - 1)} \exp(-y/2), \tag{3.16}$$

where $\Gamma(x)$ is known as the Gamma function.[8]

Figure 3.9 shows the PDFs and CDFs for three parameter values ($\nu = 1$, 3, or 5) of a χ^2 random variable. These PDFs are only

[8] The Gamma function is defined as the integral:

$$\Gamma(x) = \int_0^\infty z^{x-1} \exp(-z)dz.$$

When x is a positive integer, $\Gamma(x) = (x-1) \times (x-2) \times (x-3) \times \ldots \times 3 \times 2 \times 1 = (x-1)!$, with ! being the factorial operator.

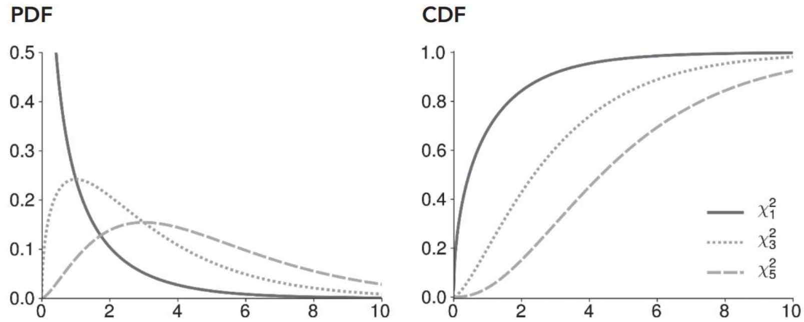


Figure 3.9 **The left panel shows the PDFs of χ^2 distributed random variables with $\nu = 1, 3,$ or 5. The right panel shows the corresponding CDFs.**

defined for positive values, because the χ^2 is the distribution of a sum of squared random variables. When $\nu = 1$, the distribution has its mode at 0. As ν increases, the density shifts to the right and spreads out, reflecting that the mean (ν) and the variance (2ν) of a χ^2_ν are both increasing with ν. In all cases, the χ^2 distribution is right-skewed. The skewness declines as the degrees of freedom increases, and when ν is large (≥ 25), a χ^2_ν random variable is well approximated by a $N(\nu, 2\nu)$.

Student's t

The Student's t distribution is closely related to the normal, but it has heavier (i.e., 'fatter') tails. The Student's t distribution was originally developed for testing hypotheses using small samples.

A Student's t is a one-parameter distribution. This parameter, denoted by ν, is also called the degrees of freedom parameter. While it affects many aspects of the distribution, the most important effect is on the shape of the tails.

A Student's t can be defined as the distribution of a standard normal random variable divided by the square root of an independent Chi-squared random variable, itself divided by its degrees of freedom

$$Y = \frac{Z}{\sqrt{W/\nu}}, \tag{3.17}$$

where Z is a standard normal, W is a χ^2_ν random variable, and Z and W are independent.[9] Dividing a standard normal by another random variable produces heavier tails than the standard normal. This is true for all values of ν, although a Student's t converges to a standard normal as $\nu \to \infty$.[10]

If $Y \sim t_\nu$, then the mean is

$$E[Y] = 0$$

and the variance is

$$V[Y] = \frac{\nu}{\nu - 2}$$

The kurtosis of Y is

$$\text{kurtosis}(Y) = 3\frac{\nu - 2}{\nu - 4}$$

The mean is only finite if $\nu > 1$ and the variance is only finite if $\nu > 2$. The kurtosis is defined for $\nu > 4$ and is always larger than 3 (i.e., it is larger than the kurtosis of a normal random variable). In general, the m^{th} moment for a t distribution will only exist for $\nu > m$.

In some applications of a Student's t, it is desirable to separate the degrees of freedom from the variance (i.e., generate a t distribution with unit variance). Using the basic result that:

$$V[aY] = a^2 V[Y]$$

it is easy to see that

$$V\left[\sqrt{\frac{\nu - 2}{\nu}}\,Y\right] = 1$$

when $Y \sim t_\nu$. This distribution is known as a standardized Student's t, because it has mean 0 and variance 1 for any value

[9] Z and W are independent if the normal random variable used to generate W is independent of Z.

[10] This should be intuitive considering that $E[W/\nu] = 1$ for all values of ν and $V[W/\nu] = 1/\nu$, so that as ν becomes larger, the denominator closely resembles a constant value.

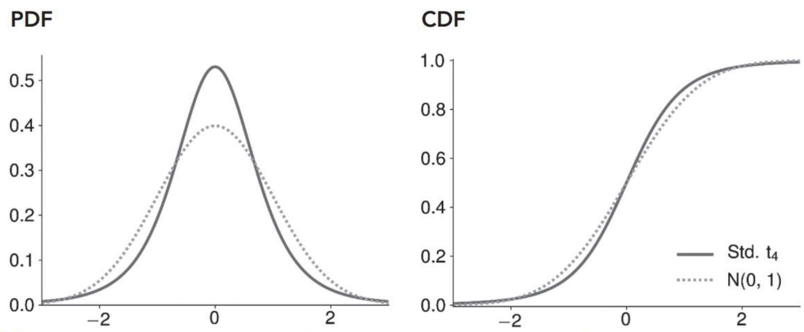


Figure 3.10 **The left panel shows the PDF of a generalized Student's *t* with four degrees of freedom and the PDF of a standard normal. The right panel shows the corresponding CDFs.**

of ν. Importantly, it can be rescaled to have any variance and re-centered to have any mean if $\nu > 2$.

Hence the generalized Student's *t* with degrees of freedom ν that has also been rescaled to have mean μ and variance σ^2 is parameterized with three parameters reflecting the mean, variance, and degrees of freedom, and is denoted as Gen. $t_\nu(\mu, \sigma^2)$.

The features of the generalized *t* make it better suited to modeling the returns of many assets than the normal distribution. It captures the heavy tails in asset returns (i.e., the increased likelihood of observing a value larger than $q\sigma$ relative to a normal for values $|q| > 2$), which the normal distribution cannot, while retaining the flexibility of the latter to directly set the mean and variance.

For example, the chance of observing a value more than 4σ away from the mean, if X_1 is normally distributed, is about 1 in 15,000. If X_2 is a standardized *t* with the same mean and variance σ with the degree of freedom parameter $\nu = 8$, then the chance of observing a value larger than 4σ is 1 in 580. This difference has important implications for risk management, especially when we are worried about rare events, and makes the standardized *t* empirically very useful.

Figure 3.10 compares the PDFs and CDFs of a generalized Student's *t* with $\nu = 4$ to a standard normal, which is the limiting distribution as $\nu \rightarrow \infty$. The Student's *t* is more peaked, heavier-tailed, and has less probability in the regions around ± 1.5. This appearance is common in heavy-tailed distributions. These two distributions have the same variance by construction, and so the increased probability of a large magnitude observation in the *t* must be offset with an increased chance of a small magnitude observation to keep the variance constant.

F

The *F* is another distribution that is commonly encountered when testing hypotheses about model parameters. The *F* has two parameters, ν_1 and ν_2, respectively known as the numerator and denominator degrees of freedom. These parameters are usually written as subscripts in the shorthand expression F_{ν_1, ν_2}.

An *F* distribution can be defined as the ratio of two independent χ^2 random variables where each has been divided by its degree of freedom

$$Y = \frac{W_1/\nu_1}{W_2/\nu_2}, \tag{3.18}$$

where $W_1 \sim \chi^2_{\nu_1}$, and $W_2 \sim \chi^2_{\nu_2}$, and W_1 and W_2 are independent.[11] If $Y \sim F_{\nu_1, \nu_2}$, then the mean of *Y* is

$$E[Y] = \frac{\nu_2}{\nu_2 - 2},$$

which is only finite when ν_2 is larger than 2. The variance of *Y* is

$$V[Y] = \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}$$

and is only finite for $\nu_2 > 4$.

When using an *F* in hypothesis testing, ν_1 is usually determined by the hypothesis being tested and is typically small (e.g., 1, 2, 3, . . .), while ν_2 is related to the sample size (and so is relatively large). When ν_2 is large, the denominator (W_2/ν_2) has mean 1 and variance[12] $2/\nu_2 \approx 0$, and so behaves like a constant. In this case,

[11] W_1 is independent of W_2 if the normal random variables used to construct W_1 are independent of the normal random variables used to construct W_2.

[12] $\approx$ means '*almost* equal to'.

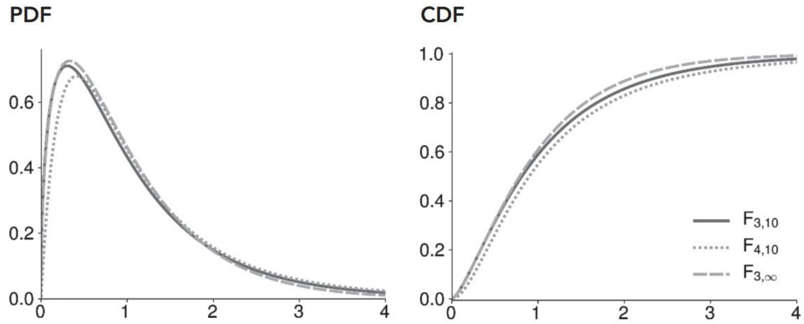


Figure 3.11 **The left panel shows the PDFs of three *F*-distributed random variables. The right panel shows the corresponding CDFs.**

if $Y \sim F_{\nu_1,\nu_2}$, then the distribution is close to W_1/ν_1, and the mean and variance are 1 and $2/\nu_1$, respectively. The PDF and CDF of an *F* are both complicated and involve integrals that do not have closed forms.

Figure 3.11 shows the PDF and CDF for three sets of parameter values of *F* distributed random variables. Comparing the $F_{3,10}$ and the $F_{4,10}$, the extra degree of freedom in the numerator (i.e., moving ν_1 from 3 to 4) has two effects. The first is that it shifts the distribution to the right. The second is that it shifts the probability closer to the mean (i.e., 1.25). The result is that these two distributions have the same mean but the $F_{3,10}$ has a larger variance. Note that values from the *F* distribution are always positive, and the PDF and CDF have roughly the same shape as the corresponding density functions for a χ^2_3.

Comparing the $F_{3,10}$ and $F_{3,\infty}$, the distribution with a larger denominator degrees of freedom parameter lies to the left and has a thinner right tail. These parameters change the mean and variance of $F_{3,\infty}$ to 1 and 2/3, respectively.

The *F* is closely related to the χ^2 and the Student's *t*. When $\nu_1 = 1$ and ν_2 is sufficiently large, then F_{1,ν_2} is close to χ^2_1. If $Y \sim t_\nu$, then $Y^2 \sim F_{1,\nu}$. These properties can be readily verified using the definitions of the *t*, χ^2 and *F* distributions.

Exponential

The exponential distribution is most commonly used to model the time until a particular event occurs. It uses a single parameter, β, that determines both the mean and variance. If $Y \sim \text{Exponential}(\beta)$, then its mean and variance are respectively

$$E[Y] = \beta$$

and

$$V[Y] = \beta^2$$

The PDF of an Exponential(β) is

$$f_Y(y) = \frac{1}{\beta} \exp\left(-\frac{y}{\beta}\right), \; y \geq 0 \tag{3.19}$$

The CDF is

$$F_Y(y) = 1 - \exp\left(-\frac{y}{\beta}\right) \tag{3.20}$$

The exponential distribution is closely related to the Poisson distribution. For example, suppose *X* is a random variable that measures the number of loan defaults per quarter. If *X* is Poisson distributed with parameter β, then the time between each subsequent loan default has as an exponential distribution with parameter β.

Exponential variables are also *memoryless*, meaning that their distributions are independent of their histories. For example, suppose that the time until a company defaults is exponentially distributed with parameter β. Assuming that β is known and constant, it follows that the probability of the company defaulting within the next year (days 0 to 365) is the same as the probability that company defaults between a year and two years from now (days 365 to 730). In fact, the probability the company defaults in any one-year window is the same (assuming it has not already defaulted). The *memoryless* property of an exponential random variable is a statement of probability conditioned on the company having not defaulted. and is based on a given length of time (which is one year in the previous example). However, it does not imply that there is equal probability today that the company defaults in in the first year (days 0 to 365) or in the first two years (days 0 to 730).

Figure 3.12 shows the PDF and CDF of three exponential distributions with β between 1 and 5. The PDF's maximum will be when *y* is zero, whatever the value of β, and is declining in *y* in

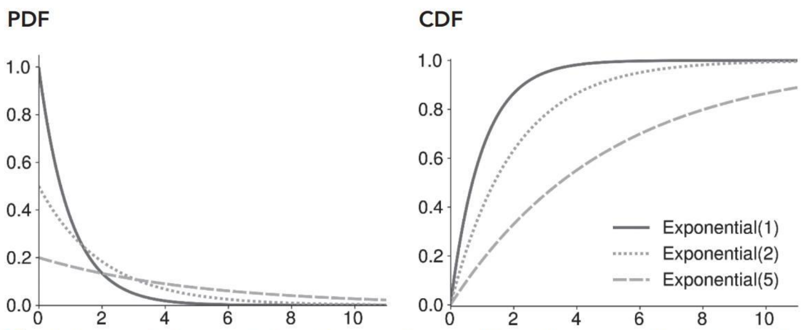


Figure 3.12 **The left panel shows the PDFs of three random variables with exponential distributions having shape parameters between 1 and 5. The right panel shows the corresponding CDFs.**

all three cases (although more quickly when β is lower). In contrast, the CDF is monotonically increasing in y in all three cases (although more quickly when β is lower as well).

As an example, assume that the time to default for a specific segment of credit card consumers is exponentially distributed with a β of five years. To find the probability that the customer will not default before year six, start by calculating the cumulative distribution until year six and then subtract this from 100%:

$$\text{Survival}_{>\text{Year 6}} = 1 - F_Y(6|\beta = 5) = 1 - \left(1 - e^{-6/5}\right) = 30.1\%$$

Beta

The Beta distribution applies to continuous random variables with outcomes between 0 and 1. It is commonly used to model probabilities that naturally fall into this range (e.g., frequent application of the beta distribution is to model the uncertainty around a probability of 'success'). The Beta distribution has two parameters, α and β, that jointly determine the mean and variance of a Beta-distributed random variable. If $Y \sim \text{Beta}(\alpha, \beta)$, its mean and variance are respectively

$$E[Y] = \frac{\alpha}{\alpha + \beta}$$

and

$$V[Y] = \frac{\alpha\beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)}$$

The PDF of a Beta(α, β) is

$$f_Y(y) = \frac{y^{(\alpha - 1)}(1 - y)^{(\beta - 1)}}{B(\alpha, \beta)}, \quad (3.21)$$

where $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha + \beta)$ and $\Gamma(\cdot)$ is the Gamma function.

$B(\alpha, \beta)$ is known as a normalization constant and it ensures that the area under the PDF curve is always one (as required for all distributions). The CDF of a Beta distribution is quite complex and its description requires functions that are beyond the scope of this text.

The left panel of Figure 3.13 shows the PDFs for four Beta distributions. Three of the examples have the same mean (i.e., 1/2). When both parameters are less than 1, the distribution places most of the probability mass near the boundaries. When $\alpha = \beta = 1$, then the Beta distribution simplifies to a standard uniform distribution. As the parameters increase, the distribution becomes more concentrated around the mean. When $\alpha = 7$ and $\beta = 3$, the distribution shifts towards the upper bound. In general, increasing α shifts the distribution towards 1, and increasing β shifts the distribution towards 0. Increasing both parameters proportionally reduces the variance and concentrates the distribution around the mean.

The Beta distribution is related to several other continuous distributions. When $\alpha = 1$ and $\beta = 1$, the Beta distribution collapses to a uniform distribution with a [0, 1] interval. When α and β are very similar and sufficiently large (e.g., $\alpha, \beta \geq 10$), the Beta distribution can be approximated by the normal.

Mixtures of Distributions

Mixture distributions build new, complex distributions using two or more component distributions. While this chapter focuses on two-component mixtures, a mixture can be constructed using any (finite) number of components.[13]

[13] Mixtures with infinitely many components are more complicated because it is not always straight forward to be assured that the density will integrate to unity. A Student's t is an infinite mixture of normal random variables where the variance of each component is constructed using draws from the χ^2_ν distribution.

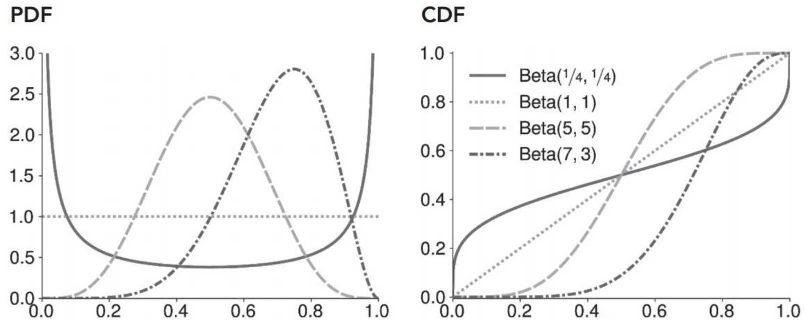


Figure 3.13 The left panel shows the PDFs for four configurations of the Beta distribution, including the case where $\alpha = \beta = 1$, which produces a standard uniform. The right panel shows the corresponding CDFs.

A two-component mixture first draws a value from a Bernoulli random variable, then, depending on the value (0 or 1), draws from one of two component distributions. This structure makes it simple to compute the CDF of the mixture when the components are normal random variables.

Suppose there is $X_1 \sim F_{X1}$, $X_2 \sim F_{X_2}$, and $W \sim$ Bernoulli(p). There mixture of X_1 and X_2 would therefore be

$$Y = WX_1 + (1 - W)X_2$$

If we sampled from this distribution, $100 \times p\%$ of draws would come from the distribution of X_1, and the remainder would be from the distribution of X_2.

Both the CDF and the PDF of a mixture distribution are the weighted averages of the CDFs or PDFs (respectively) of the components

$$F_Y(y) = pF_{X_1}(y) + (1 - p)F_{X_2}(y) \quad (3.22)$$

$$f_Y(y) = pf_{X_1}(y) + (1 - p)f_{X_2}(y) \quad (3.23)$$

While directly computing the central moments of a mixture distribution is challenging, doing so for non-central moments is simple. For example, the mean of the mixture is:

$$E[Y] = pE[X_1] + (1 - p)E[X_2] \quad (3.24)$$

and the second non-central moment is

$$E[Y^2] = pE[X_1^2] + (1 - p)E[X_2^2] \quad (3.25)$$

These two non-central moments can be combined to construct the variance

$$V[Y] = E[Y^2] - E[Y]^2$$

Higher-order, non-central moments are equally easy to compute because they are weighted versions of the non-central moments of the mixture's components. Using the relationship between the third and fourth central and non-central moments defined in Chapter 2, these central moments are related to the non-central moments by

$$E[(Y - E[Y])^3] = E[Y^3] - 3E[Y^2]E[Y] + 2E[Y]^3$$
$$E[(Y - E[Y])^4] = E[Y^4] - 4E[Y^3]E[Y] + 6E[Y^2]E[Y]^2 - 3E[Y]^4 \quad (3.26)$$

An important feature of mixtures is that they can have both skewness and excess kurtosis even when their components have no skewness or excess kurtosis (e.g., when the components are normal random variables).

For example, consider a distribution generated by mixing two normal random variables:

$$Y = WX_1 + (1 - W)X_2,$$

where $X_1 \sim N(-0.9, 1 - 0.9^2)$, $X_2 \sim N(0.9, 1 - 0.9^2)$, and $W \sim$ Bernoulli($p = 0.5$). The parameters of the three component normal distributions and the Bernoulli have been chosen so that $E[Y] = 0$ and $V[Y] = 1$.

In this case, the mixing probability is 50%. The mean of Y is 0 by construction because

$$\begin{aligned} E[Y] &= E[W X_1 + (1 - W)X_2] \\ &= E[W X_1 + X_2 - W X_2] \\ &= E[W]E[X_1] + E[X_2] - E[W]E[X_2] \\ &= 0.5(-0.9) + 0.9 - 0.5(0.9) \\ &= 0 \end{aligned}$$

Similarly, the variance is 1 because

$$E[X_1^2] = E[X_2^2] = E[Y^2] = 0.9^2 + (1 - 0.9^2) = 1$$

and thus

$$V[Y] = E[Y^2] - E[Y]^2 = 1 - 0 = 1$$

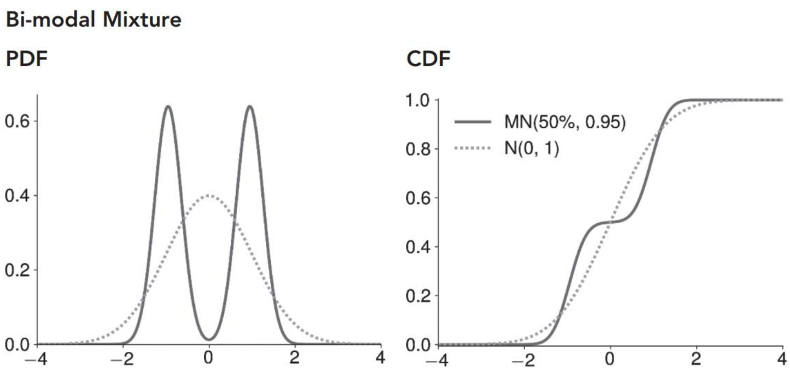


Figure 3.14 **The panels show the PDFs (left) and CDFs (right) of a bimodal mixture where each component has probability 50% of occurring and is a standard normal (N(0, 1)).**

As seen in Figure 3.14, the mixture build from these two components is bimodal. This is because each component has a small variance and the two means are distinct.

As another example, consider a mixture of $X_1 \sim N(0, 0.725^2)$ and $X_2 \sim N(0, 3.16^2)$, weighted 0.95 and 0.05 (respectively). Note that X_1 is close to a standard normal, whereas X_2 has a much greater variance but also a much lower probability of occurrence. As such, draws from X_2 appear as outliers in the mixture distribution.

The result is a mixture distribution that is very heavy-tailed with a kurtosis of 15.7. Figure 3.15 compares the PDF and the CDF of a contaminated normal with that of a standard normal. Note that the contaminated normal is similar in shape to a generalized Student's *t*, as it is both more peaked and heavier-tailed then the normal. The heavy tails can be seen in the CDF by the crossing of the two curves for values below −2 and above 2, whereas the normal CDF is closer to the boundary (i.e., 0 or 1) than the CDF of the contaminated normal.

Mixing components with different means and variances produces a distribution that is both skewed and heavy-tailed. For example, consider a mixture with components $X_1 \sim N(0.16, 0.61)$ and

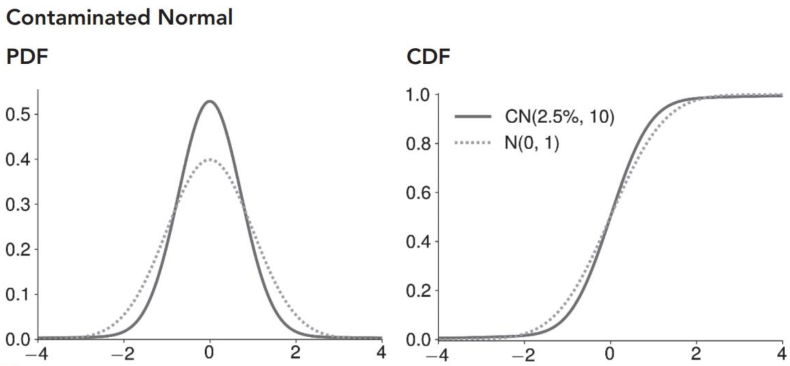


Figure 3.15 **This shows the PDFs (left) and CDFs (right) for a contaminated normal (CN(2.5%, 10)) and a standard normal.**

Mixed Normal

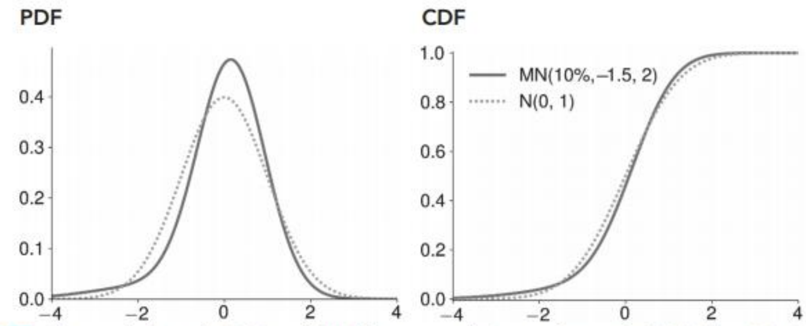


Figure 3.16 The panels show the PDFs and CDFs for a general mixture of normals ($MN(10\%, -1.5, 2)$) and a standard normal.

$X_2 \sim N(-1.5, 2)$, where the probability of drawing from the distribution of X_1 is $p = 90\%$. The mean and variance of this mixture are 0 and 1, respectively. The skewness and kurtosis are -0.96 and 5.50, respectively. As shown in Figure 3.16, this distribution has a strong left skew and is mildly heavy-tailed. Specifically, the mixture is clearly left-skewed with much more probability between -4 and -2 than between 2 and 4.

3.3 SUMMARY

This chapter introduces the key distributions—both discrete and continuous—commonly encountered in finance and risk management. Familiarity with these distributions is important when building financial data models because each distribution has a specific feature set that makes it applicable to data with corresponding features.

The simplest is the Bernoulli, which can be applied to any binary event. The binomial generalizes the Bernoulli to the distribution of n independent events. The simplest continuous distribution is the uniform, which specifies that all values within its support are equally likely. The normal is the most widely used distribution and has a broad range of applications, (e.g., modeling financial returns). The normal is also widely used when testing hypotheses about unknown parameters.

The lognormal distribution, which is a simple transformation of a normal distribution, is often assumed to describe price processes and is the distribution underlying the famous Black-Scholes Merton model. Three common continuous distributions, the Student's t, the χ^2, and the F, are all transformations of independent normal random variables. The χ^2 and the F are also widely used in hypotheses testing, while the Student's t can be used to model returns that are heavier-tailed than normal random variables.

The chapter concludes by introducing mixture distributions. This class of distributions allows multiple simple distributions to be combined to produce distributions with empirically important features (e.g., skewness and heavy tails).

The following questions are intended to help candidates understand the material. They are *not* actual FRM exam questions.

QUESTIONS

Short Concept Questions

3.1 What is the name of the distribution that describes any random binary event?

3.2 Under what conditions can a binomial and a Poisson be approximated by a normal random variable?

3.3 What type of events are Poisson random variables used to describe?

3.4 If X is a standard uniform, what is $\Pr(0.2 < Z < 0.4)$? What about $\Pr(l < X < u)$, where $0 < l < u < 1$?

3.5 How much probability is within $\pm 2\sigma$ of the mean of a normal random variable?

3.6 How many independent standard normal random variables are required to construct a χ^2_ν?

3.7 If a Beta-distributed random variable is used to describe some data, what requirement must the data satisfy?

3.8 How are the mean and variance of a mixture of normal random variables related to the mean and variance of the components of the mixture?

Practice Questions

3.9 Either using a Z table or the Excel function NORM.S.DIST, compute:

a. $\Pr(-1.5 < Z < 0)$, where $Z \sim N(0, 1)$
b. $\Pr(Z < -1.5)$, where $Z \sim N(0, 1)$
c. $\Pr(-1.5 < X < 0)$, where $X \sim N(1, 2)$
d. $\Pr(X > 2)$, where $X \sim N(1, 2)$
e. $\Pr(W > 12)$, where $W \sim N(3, 9)$

3.10 Either using a Z table or the Excel function NORM.S.INV, compute

a. z so that $\Pr(z < Z) = 0.95$ when $Z \sim N(0, 1)$
b. z so that $\Pr(z > Z) = 0.95$ when $Z \sim N(0, 1)$
c. z so that $\Pr(-z < Z < z) = 0.75$ when $Z \sim N(0, 1)$
d. a and b so that $\Pr(a < X < b) = 0.75$ and $\Pr(a < X) = 0.125$ when $X \sim N(2, 4)$

3.11 If the return on a stock, R, is normally distributed with a daily mean of 8%/252 and a daily variance of $(20\%)^2/252$, find the values where:

a. $\Pr(R < r) = 0.001$
b. $\Pr(R < r) = 0.01$
c. $\Pr(R < r) = 0.05$

3.12 The monthly return on a hedge fund portfolio with USD 1 billion in assets is $N(0.02, 0.0003)$. What is the distribution of the gain in a month?

a. The fund has access to a USD 10 million line of credit that does not count as part of its portfolio. What is the chance that the firm's loss in a month exceeds this line of credit?
b. What would the line of credit need to be to ensure that the firm's loss was less than the line of credit in 99.9% of months (or equivalently, larger than the LOC in 0.1% of months)?

3.13 If the kurtosis of some returns on a small-cap stock portfolio was 6, what would the degrees of freedom parameter be if they were generated by a generalized Student's t_ν? What if the kurtosis was 9?

3.14 An analyst is using the following exponential function to model corporate default rates:

$$f_Y(y) = \frac{1}{\beta} \exp\left(-\frac{y}{\beta}\right),\ y \geq 0,$$

where y is the number of years.

a. What is the cumulative probability of default within the first five years?
b. What is the cumulative probability of default within the first ten years given that the company has survived for five years?

The following questions are intended to help candidates understand the material. They are *not* actual FRM exam questions.

ANSWERS

Short Concept Questions

3.1 Any binary random variable can be described as a Bernoulli by mapping the outcomes to 0 or 1. For example, a corporate default (1 for default) or the direction of the market (1 for positive return).

3.2 A binomial can be well approximated by a normal when np and $n(1 - p)$ are both greater than or equal to 10. A Poisson can be accurately approximated by a normal when the parameter λ is large. The approximation will become more accurate as λ increases. The standard recommendation for approximating the Poisson with a normal requires $\lambda \geq 1000$.

3.3 Poisson random variables are used for counts—the number of events in a fixed unit of time. For example, the number of mortgage defaults over a month.

3.4 In a standard uniform, the probability between two points only depends on the distance between them, so $\Pr(0.2 < Z < 0.4) = 0.4 - 0.2 = 20\%$ and $\Pr(l < X < u) = u - l$.

3.5 Formally this quantity is 95.45%, but this area is commonly approximated as 95%.

3.6 A χ^2_ν is defined as the sum of ν in dependent standard normal random variables, and so ν.

3.7 Beta random variables are continuous values on the interval [0,1], and so the values must lie in this range.

3.8 The mean is the weighted average of the means of the components. The variance is more complicated. The second *non-central* moment, $E[X^2]$ of the mixture is the weighted average of the second non-central moments of the components. The variance is then $E[X^2] - E[X]^2$, which depends the first and second moments of the mixture.

Solved Problems

3.9 **a.** 43.3%. In Excel, the command to compute this value is NORM.S.DIST(0,TRUE) – NORM.S.DIST(−1.5,TRUE).

b. 6.7%. In Excel, the command to compute this value is NORM.S.DIST(−1.5,TRUE).

c. 20.1%. In Excel, the command to compute this value is NORM.S.DIST((0 – 1)/SQRT(2),TRUE) – NORM.S.DIST((−1.5−1)/SQRT(2),TRUE).

d. 24.0%. In Excel, the command to compute this value is 1 – NORM.S.DIST((2 – 1)/SQRT(2),TRUE).

e. 0.13%. In Excel, the command to compute this value is 1 – NORM.S.DIST((12 – 3)/3,TRUE).

3.10 **a.** 1.645. In Excel, the command to compute this value is NORM.S.INV(0.95).

b. −1.645. In Excel, the command to compute this value is NORM.S.INV(0.05).

c. 1.15. Here the tail to the left should have 12.5% and the tail to the right should also have 12.5%. In Excel, the command to compute this value is –NORM.S.INV(0.125).

d. −0.3 and 4.3. The area of the left and right should each have 12.5%. These can be constructed using the answer to the previous problem by re-centering on the mean and scaling by the standard deviation, so that $a = 2 \times -1.15 + 2$ and $b = 2 \times 1.15 + 2$. Note that the formula is $a = \sigma \times q + \mu$, where q is the quantile value.

3.11 **a.** The mean is 0.031% per day and the variance is 1.58% per day (so that the standard deviation is 1.26% per day). To find these values, we transform the variable to be standard normal, so that

$$\Pr(R < r) = 0.001 = \Pr\left(Z < \frac{r - \mu}{\sigma}\right) = 0.001.$$

The value for the standard normal is −3.09 (NORM.S.INV(0.001) in Excel) so that $-3.09 \times \sigma + \mu = -3.86\%$.

b. The same idea can be used here where $z = -2.32$ so that $\Pr(Z < z) = 0.01$. Transforming this value, $r = -2.32 \times \sigma + \mu = -2.89\%$.

c. Here the value of z is −1.645 so that $r = -1.645 \times \sigma + \mu = -2.04\%$. These are all common *VaR* quantiles and suggest that there is a 5% chance that the return would be less than 2.04% on any given day, a 1% chance that it would be less than 2.89%, and a

one in 1,000 chance that the return would be less than 3.86%, if returns were normally distributed.

3.12 The monthly return is 2% and the monthly standard deviation is 1.73%. The monthly change in portfolio value will also be normally distributed with a mean of $0.02 \times$ USD 1 billion = USD 20 million and a variance of $0.0003 \times 1\ \text{billion}^2 = 300$ trillion. The standard deviation of the gain will be USD 17.3 million and is found by taking the square root of the USD300 trillion variance.

a. The probability that the portfolio loses more than USD 10 million is then (working in millions):

$$\Pr(V < -10) = \Pr\left(\frac{V - 20}{17.3} < \frac{-10 - 20}{17.3}\right) = \Pr(Z < -1.73)$$

Using the normal table, $\Pr(Z < -1.73) = 4.18\%$.

b. Here we work in the other direction. First, we find the quantile where $\Pr(Z < z) = 99.9\%$, which gives $z = -3.09$. This is then scaled to the distribution of the change in the value of the portfolio by multiplying by the standard deviation and adding the mean, $17.3 \times -3.09 + 20 = -33.46$. The fund would need a line of credit of USD 33.46 million to have a 99.9% chance of having loss below this level.

3.13 In a Student's t, the kurtosis depends only on the degree of freedom and is $\kappa = 3\frac{(\nu-2)}{\nu-4}$. This can be solved so that $\kappa(\nu - 4) = 3(\nu - 2)$ so that $\kappa\nu - 4\kappa = 3\nu - 6$ and $\kappa\nu - 3\nu = 4\kappa - 6$. Finally, solving for ν, $\nu = \frac{4\kappa-6}{\kappa-3}$. Plugging in $k = 6$ gives $\nu = \frac{24-6}{6-3} = \frac{18}{3} = 6$ and plugging in $k = 9$ gives $\nu = \frac{36-6}{9-3} = 5$. The kurtosis falls rapidly as ν grows. For example, if $\nu = 12$ then $\kappa = 3.75$, which is only slightly higher than the kurtosis of a normal (3).

3.14 **a.** $F_Y(5) = 1 - \exp\left(-{}^{5}/_{\beta}\right)$

b. We need to divide the marginal probability of default between years five and ten:

$$F_Y(10) - F_Y(5)$$

By the SURVIVAL probability through year five (1 – result from part a).

$$F_Y(10) - F_Y(5) = \exp\left(-{}^{5}/_{\beta}\right) - \exp\left(-{}^{10}/_{\beta}\right)$$

$$\frac{F_Y(10) - F_Y(5)}{1 - F_Y(5)} = \frac{\exp\left(-{}^{5}/_{\beta}\right) - \exp\left(-{}^{10}/_{\beta}\right)}{\exp\left(-{}^{5}/_{\beta}\right)} = 1 - \exp\left(-{}^{5}/_{\beta}\right)$$

4

Multivariate Random Variables

Learning Objectives

After completing this reading you should be able to:

- Explain how a probability matrix can be used to express a probability mass function.
- Compute the marginal and conditional distributions of a discrete bivariate random variable.
- Explain how the expectation of a function is computed for a bivariate discrete random variable.
- Define covariance and explain what it measures.
- Explain the relationship between the covariance and correlation of two random variables and how these are related to the independence of the two variables.
- Explain the effects of applying linear transformations on the covariance and correlation between two random variables.
- Compute the variance of a weighted sum of two random variables.
- Compute the conditional expectation of a component of a bivariate random variable.
- Describe the features of an independent and identically distributed (iid) sequence of random variables.
- Explain how the iid property is helpful in computing the mean and variance of a sum of iid random variables.

梵辞学府

Multivariate random variables extend the concept of a single random variable to include measures of dependence between two or more random variables. Note that all results from Chapter 2 are directly applicable here because each component of a multivariate random variable is a univariate random variable. This chapter focuses on the extensions required to understand multivariate random variables, including how expectations change, new moments, and additional characterizations of uncertainty. While this chapter discusses discrete and continuous multivariate random variables in separate sections, note that virtually all results for continuous variables can be derived from those for discrete variables by replacing the sum with an integral.

This chapter focuses primarily on bivariate random variables to simplify the mathematics required to understand the key concepts but all definitions and results extend directly to random variables with three or more components.

4.1 DISCRETE RANDOM VARIABLES

Multivate random variables are vectors[1] of random variables. For example, a bivariate random variable *X* would be a vector with two components: X_1 and X_2. Similarly, its realization (i.e., *x*) would have two component values as well: x_1 and x_2. Treated separately, x_1 is a realization from X_1 and x_2 is a realization from X_2.

Multivariate random variables are like their univariate counterparts in that they can be discrete or continuous. Similarly, both types of random variables are denoted with uppercase letters (e.g., *X*, *Y*, or *Z*).

Note that the probability mass function (PMF)/probability density function (PDF) for a bivariate random variable returns the probability that *two* random variables each take a certain value. This means that plotting these functions requires three axes: X_1, X_2, and the probability mass/density. Because the CDF is either the sum of the PMF for all values in the support that are less than or equal to *x* (for discrete random variables) or the integral of the PDF (for continuous random variables), it requires three axes as well. The PMF/PDF and CDF will therefore be represented as surfaces on three-dimensional plots.

Probability Matrices

The PMF of a bivariate random variable is a function that returns the probability that $X = x$. Put in terms of the components, this would mean $X_1 = x_1$ and $X_2 = x_2$

$$f_{X_1,X_2}(x_1, x_2) = \Pr(X_1 = x_1, X_2 = x_2) \quad (4.1)$$

[1] A vector of dimension *n* is an ordered collection of *n* elements, which are called components.

A PMF describes the probability of the outcomes as a function of the coordinates x_1 and x_2. The probabilities are always non-negative, less than or equal to 1, and the sum across all possible values in the support of X_1 and X_2 is 1.

The leading example of a discrete bivariate random variable is the trinomial distribution,[2] which is the distribution of *n* independent trials where each trial produces one of three outcomes. In other words, this distribution generalizes the binomial.

The two components of a trinomial are X_1 and X_2, which count the number of realizations of outcomes 1 and 2. The count of the realizations for outcome 3 (i.e., X_3) is given by $n - X_1 - X_2$, and so it is redundant given knowledge of X_1 and X_2.

For example, consider the credit quality of a diversified bond portfolio. The *n* bonds in the portfolio can be classified as investment grade, high yield, or unrated. In this case, X_1 is the count of the investment grade bonds and X_2 is the count of the high yield bonds.

The trinomial PMF has three parameters:

1. *n* (i.e., the total number of experiments),
2. p_1, (i.e., the probability of observing outcome 1), and
3. p_2 (i.e., the probability of observing outcome 2).

Because these outcomes are exclusive, the probability of outcome 3 is simply

$$p_3 = 1 - p_2 - p_1$$

The PMF of a trinomial random variable is

$$f_{X_1,X_2}(x_1, x_2) = \frac{n!}{x_1!x_2!(n - x_1 - x_2)!} p_1^{x_1} p_2^{x_2} (1 - p_1 - p_2)^{n - x_1 - x_2} \quad (4.2)$$

Figure 4.1 plots the PMF of a trinomial with $p_1 = 20\%$, $p_2 = 50\%$ and $n = 5$. The mass function shows that most of the mass occurs when X_1 is small (i.e., 0 or 1).

The CDF of a bivariate variable is a function returning the total probability that each component is less than or equal to a given value for that component, so that:

$$F_{X_1,X_2}(x_1, x_2) = \sum_{\substack{t_1 \in R(X_1) \\ t_1 \leq x_1}} \sum_{\substack{t_2 \in R(X_2) \\ t_2 \leq x_2}} f_{X_1,X_2}(t_1, t_2) \quad (4.3)$$

In the equation above, t_1 contains the values that X_1 may take as long as $t_1 \leq x_1$, and t_2 is similarly defined for X_2.

Discrete distributions defined over a finite set of values can be described using a probability matrix, which relates realizations

[2] Both the binomial and the trinomial are special cases of the multinomial distribution, which is the distribution of *n* independent trials that can take one of *k* possible outcomes.

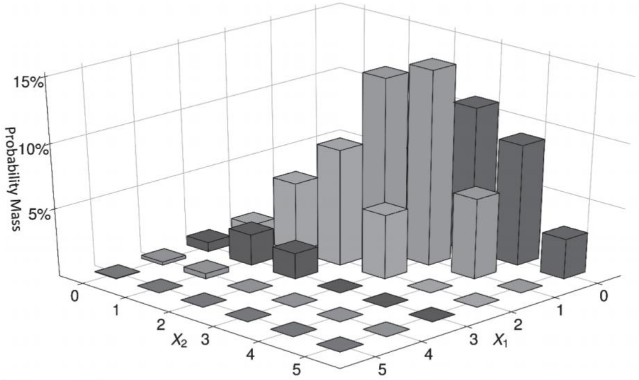


Figure 4.1 The PMF of a trinomial random variable with $p_1 = 20\%$, $p_2 = 50\%$ and $n = 5$.

to probabilities. In other words, it is a tabular representation of a PMF.

For example, suppose that the return on a company's stock is related to the rating given to the company by an analyst. For simplicity, assume that the stock can only have one of three possible returns: 0, 5%, or −5%. Analyst ratings can be positive, neutral, or negative, and are labeled −1, 0, or 1 (respectively). The probability matrix for this problem is shown in Table 4.1.

Each cell contains the probability that the combination of the two outcomes is realized. For example, there is a 10% chance that the stock price declines 5% with a neutral rating and a 20% chance that it declines 5% with a negative rating.

Marginal Distributions

The PMF describes the *joint* distribution of the two components [i.e., $\Pr(X_1 = x_1, X_2 = x_2)$] and provides a complete description of the uncertainty across both random variables. However, it is also possible to examine each variable individually.

Table 4.1 Probability Matrix for Analyst Ratings and Stock Returns

	Analyst Ratings		Stock Return (X_1)		
Analyst (X_2)			−5%	0%	5%
	Negative	−1	20%	10%	0%
	Neutral	0	10%	15%	15%
	Positive	1	5%	5%	20%

The distribution of a single component of a bivariate random variable is called a marginal distribution, and it is simply a univariate random variable. The focus of the chapter is on bivariate random variables, and so a marginal distribution can only be defined for a single random variable. For a random variable with k-components, the marginal distribution can be computed for any subset of the component random variable containing between 1 and $k - 1$ components. Any marginal distribution is constructed from the joint distribution by summing the probability over the *excluded* components.

The marginal distribution of X_1 contains the probabilities of realizations of X_1 and its PMF denoted by $f_{X_1}(x_1)$. Note that this is the same notation used to describe the PMF of a univariate random variable in Chapter 2.

The marginal PMF of X_1 can be computed by summing the probability for each realization of X_1 across all values in the support of X_2. The marginal PMF is defined as:

$$f_{X_1}(x_1) = \sum_{x_2 \in R(X_2)} f_{X_1,X_2}(x_1, x_2) \qquad (4.4)$$

Returning to the previous example using the probability matrix, the marginal PMF of the recommendation when the stock return is 5% is:

$$f_{X_1}(5\%) = \sum_{x_2=\{-1,0,1\}} f(5\%, x_2) = 0\% + 15\% + 20\% = 35\%$$

This calculation can be repeated for stock returns of −5% and 0% to construct the complete marginal PMF for each of the three possible stock returns, which is

	Marginal PMF		
Stock Return	−5%	0%	5%
Probability	35%	30%	35%

When a PMF is represented as a probability matrix, the two marginal distributions are computed by summing across columns (which constructs the marginal distribution of the row variables) or summing down rows (which constructs the marginal PMF for the column variables).

The two marginal distributions in the final stock return/recommendation example are given in the final column and row of Table 4.2.

Table 4.2 **Probability Matrix with Marginal Distributions for Analyst Recommendations and Stock Returns**

			Stock Return (X_1)			
			−5%	0%	5%	$f_{X_2}(x_2)$
Analyst (X_2)	Negative	−1	20%	10%	0%	30%
	Neutral	0	10%	15%	15%	40%
	Positive	1	5%	5%	20%	30%
		$f_{X_1}(x_1)$	35%	30%	35%	

Note that the marginal PMF is a univariate PMF.

The interpretation of the marginal probabilities is straightforward. For example, ignoring the stock returns, the probability of a negative analyst rating is 30%; ignoring the analyst ratings, the probability of a negative return is 35%.

The marginal CDF is defined in a natural way using the marginal PMF so that $\Pr(X_1 < x_1)$ measures the total probability in the marginal PMF less than x_1:

$$F_{X_1}(x_1) = \sum_{\substack{t_1 \in R(X_1) \\ t_1 \leq x_1}} f_{X_1}(t_1)$$

Independence

In Chapter 1, two events *A* and *B* were defined to be independent if:

$$\Pr(A \cap B) = \Pr(A)\Pr(B)$$

This definition of independence extends directly to bivariate random variables, and the components of a bivariate random variable are independent if:

$$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2) \quad (4.5)$$

Independence requires that the joint PMF of X_1 and X_2 be the product of the marginal PMFs.

Returning to the example of the stock return and ratings, the original joint distribution (left) and the product of the marginals (right) are shown in Table 4.3.

Table 4.3 **Joint Distribution and Products of Marginals for Analyst Ratings and Stock Returns**

			Joint Distribution			Product of Marginals		
			Stock Return (X_1)			Stock Return (X_1)		
			−5%	0%	5%	−5%	0%	5%
Analyst (X_2)	Negative	−1	20%	10%	0%	10.5%	9%	10.5%
	Neutral	0	10%	15%	15%	14%	12%	14%
	Positive	1	5%	5%	20%	10.5%	9%	10.5%

The matrix on the left is constructed by multiplying the marginal PMF of the stock return and the marginal PMF of the analyst rating. For example, the marginal probability that the stock return is −5% is the sum of the first column (i.e., 35%). The probability that the analyst gives a negative rating is the sum of the first row (i.e., 30%). The upper-left value in the right panel is then

$$35\% \times 30\% = 10.5\%.$$

This step can be repeated to obtain all nine elements of the products of the marginals table on the right. If two random variables are independent, one must contain no information about the other. However, this condition is violated here because if we know that if $X_1 = 5\%$, then X_2 cannot be negative. Therefore, the two random variables are dependent (i.e., not independent).

An alternative way to confirm that the analyst ratings and stock returns are not independent is to compare the elements of the two tables. Since none of the elements of the joint distribution are equal to their corresponding values from the product of the marginals table, condition (4.5) cannot hold, and therefore X_1 and X_2 cannot be independent.

Conditional Distributions

Marginal PMFs summarize the probabilities of the outcomes for each component and are thus univariate PMFs. The conditional distribution, on the other hand, summarizes the probability of the outcomes for one random variable *conditional* on the other taking a specific value. In Chapter 1, the conditional probability of two events was defined as:

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

This definition of conditional probability can be applied to a bivariate random variable to construct a conditional distribution. The conditional distribution of X_1 given X_2 is defined as:

$$f_{X_1|X_2}(x_1|X_2 = x_2) = \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)} \quad (4.6)$$

In other words, it is the joint probability of the two events divided by the marginal probability that $X_2 = x_2$. For example, suppose that we want to determine the distribution of stock returns (X_1) conditional on a positive analyst rating ($X_2 = 1$). This conditional distribution is:

$$f_{X_1|X_2}(x_1|X_2 = 1) = \frac{f_{X_1,X_2}(x_1, 1)}{f_{X_2}(1)} = \frac{f_{X_1,X_2}(x_1, 1)}{0.3}$$

The value of $f_{X_2}(1)$ in the denominator comes from the marginal distribution for analyst recommendations in the final column of the Table 4.2. Applying this expression to the row corresponding to $X_2 = 1$ in the bivariate probability matrix, the conditional distribution is then:

	Marginal PMF		
Stock Return	−5%	0%	5%
Probability	16.6%	16.6%	66.6%

Recall that probabilities of the three outcomes of the stock return are 5%, 5%, and 20% when the analyst makes a positive recommendation. These values are divided by the marginal probability of a positive recommendation (i.e., 30%) to produce the conditional distribution. Note that a conditional PMF must satisfy the conditions of any PMF, and so the probabilities are always non-negative and sum to one.

Conditional distributions can also be defined over a set of outcomes for one of the variables. For example, consider the distribution of the stock return given that the analyst did *not* give a positive rating. This set is:

$$X_2 \in \{-1,0\}$$

The conditional PMF must sum across all outcomes in the set that is conditioned on $S = \{-1,0\}$, and so:

$$f_{X_1|X_2}(x_1|X_2 \in S) = \frac{\Sigma_{x_2 \in S} f_{X_1,X_2}(x_1, x_2)}{\Sigma_{x_2 \in S} f_{X_2}(x_2)} \quad (4.7)$$

The marginal probability that $X_2 \in \{-1,0\}$ is the sum of the (marginal) probabilities of these two outcomes, again from the final column of Table 4.2:

$$f_{X_2}(-1) + f_{X_2}(0) = 30\% + 40\% = 70\%$$

The conditional probability is then:

$$f_{X_1|X_2}(x_1|X_2 \in \{-1,0\}) = \begin{cases} \frac{20\% + 10\%}{70\%} = 42.8\% \\ \frac{10\% + 15\%}{70\%} = 35.7\% \\ \frac{0\% + 15\%}{70\%} = 21.4\% \end{cases}$$

The numerators are simply the sums of the probabilities of the stock return taking a given value (i.e., −5%, 0%, or 5%) for the neutral or negative ratings.

The definition of a conditional distribution can also be rearranged so that the joint PMF is expressed in terms of the marginal PMF and a conditional PMF. For example:

$$f_{X_1,X_2}(x_1, x_2) = f_{X_1|X_2}(x_1|X_2 = x_2)f_{X_2}(x_2)$$

or

$$f_{X_1,X_2}(x_1, x_2) = f_{X_2|X_1}(x_2|X_1 = x_1)f_{X_1}(x_1)$$

These identities can be transformed to produce further insights. First, recall that the components of a bivariate random variable are independent if:

$$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1)f_{X_2}(x_2)$$

Comparing this definition to the decomposition of the joint into the conditional and the marginal, then independence requires that:

$$f_{X_1|X_2}(x_1|X_2 = x_2) = f_{X_1}(x_1)$$

In other words, equality of the conditional and the marginal PMFs is a requirement of independence. Specifically, knowledge about the value of X_2 (e.g., $X_2 = x_2$) must not contain any information about X_1.

Returning to the example of the stock return and recommendation, the marginal distribution of the stock return and the conditional distribution when the analyst gives a positive rating are

	Marginal PMF			Conditional PMF on $X_2 = 1$		
Stock Return	−5%	0%	5%	−5%	0%	5%
Probability	35%	30%	35%	16.6%	16.6%	66.6%

These distributions differ, and, therefore, reconfirm that the recommendation and the return are not independent.

4.2 EXPECTATIONS AND MOMENTS

Expectations

The expectation of a function of a bivariate random variable is defined analogously to that of a univariate random variable.[3] The expectation of a function $g(X_1, X_2)$ is a probability weighted average of the function of the outcomes $g(x_1, x_2)$. The expectation is defined as:

$$E[g(X_1, X_2)] = \sum_{x_1 \in R(X_1)} \sum_{x_2 \in R(X_2)} g(x_1, x_2) f_{X_1,X_2}(x_1, x_2) \quad (4.8)$$

Note that while $g(x_1, x_2)$ depends on x_1 and x_2, it may be a function of only one of the components. The double sum across the support of both X_1 and X_2 is required when computing any expectation—even when the function depends on only one component.

To be clear, $g(\cdot)$ are the outcomes and $f(\cdot)$ are the probabilities associated with these outcomes. The expectation is simply the sum of each possible outcome across dimensions 1 and 2 weighted by its probability of occurrence. As in the univariate case, $E[g(X_1, X_2)] \neq g(E[X_1],E[X_2])$ for a nonlinear function $g(x_1, x_2)$.

As an example of how to calculate the expectation in the context of a non-linear function, consider the following joint PMF:

[3] As a reminder, the expectation is the weighted sum of the outcome times the probability of the outcome.

		X_1	
		1	2
X_2	3	15%	10%
	4	60%	15%

In this case, variable X_1 can take the values 1 or 2 whereas X_2 can take the values 3 or 4. The joint probability that $X_1 = 1$ and $X_2 = 3$ is 15%, and so on.

Now, consider the function

$$g(x_1, x_2) = x_1^{x_2}$$

The expectation of $g(x_1, x_2)$ is therefore

$$E[g(x_1, x_2)] = \sum_{x_1 \in \{1,2\}} \sum_{x_2 \in \{3,4\}} g(x_1, x_2) f(x_1, x_2)$$

$$\begin{aligned} E[g(x_1, x_2)] &= 1^3(0.15) + 1^4(0.6) + 2^3(0.10) + 2^4(0.15) \\ &= 0.15 + 0.60 + 0.80 + 2.4 \\ &= 3.95 \end{aligned}$$

Notice that the double sum has four elements, since there are four possible outcomes, each weighted by their probability.

Moments

Expectations are used to define moments of bivariate random variables in the same way that they are used to define moments for univariate random variables. Since $X = [X_1, X_2]$, the first moment of X (i.e., the mean E[X]) is the mean of the components

$$E[X] = [E[X_1], E[X_2]] = [\mu_1, \mu_2] \quad (4.9)$$

The second moment of X adds an extra term and is referred to as the covariance of X. The covariance is technically a 2-by-2 matrix of values, where the values along one diagonal are the variances of X_1 and X_2, and the terms in the other diagonal are the covariance between X_1 and X_2. More specifically, the variances are on the diagonal running from top left to bottom right in the matrix, which is known as the *leading diagonal*.

The variance is a measure of dispersion for a single variable, and the definition of the variance of each component is unchanged:

$$V[X_1] = E[(X_1 - E[X_1])^2] \quad (4.10)$$

The covariance between X_1 and X_2 is defined as:

$$\begin{aligned} Cov[X_1, X_2] &= E[(X_1 - E[X_1])(X_2 - E[X_2])] \\ &= E[X_1 X_2] - E[X_1]E[X_2] \quad (4.11) \end{aligned}$$

The covariance is a measure of dispersion that captures how the variables move together. Note that it is a generalization of the variance, and the covariance of a variable with itself is just the variance:

$$\begin{aligned} Cov[X_1, X_1] &= E[(X_1 - E[X_1])(X_1 - E[X_1])] \\ &= E[(X_1 - E[X_1])^2] \\ &= V[X_1] \end{aligned}$$

In a bivariate random variable, there are two variances and one covariance since $Cov[X_1, X_2] = Cov[X_2, X_1]$ and thus the covariance matrix will be symmetric about the leading diagonal. Therefore, it is necessary to distinguish between these when using short-hand notation. It is common to express $V[X_1]$ as σ_1^2 and $V[X_2]$ as σ_2^2. The symbols σ_1 and σ_2 are used to denote the standard deviations of X_1 and X_2, respectively.

An alternative abbreviation scheme uses σ_{11} and σ_{22} to indicate the variances of X_1 and X_2, respectively. The short-hand notation for their standard deviations is identical in this alternative scheme. The covariance is commonly abbreviated to σ_{12} in both schemes.

The covariance depends on the scales of X_1 and of X_2. It is therefore more common to report the correlation, which is a scale-free measure. The correlation is defined as:

$$Corr[X_1, X_2] = \frac{Cov[X_1, X_2]}{\sqrt{V[X_1]}\sqrt{V[X_2]}} = \frac{\sigma_{12}}{\sqrt{\sigma_1^2}\sqrt{\sigma_2^2}} = \frac{\sigma_{12}}{\sigma_1 \sigma_2} \quad (4.12)$$

Correlation measures the strength of the linear relationship between two variables and is always between −1 and 1. For example, if $X_2 = a + bX_1$, then the correlation between X_1 and X_2 is 1 if $b > 0$, −1 if $b < 0$ or 0 if $b = 0$. This can be directly verified because:

$$\begin{aligned} Cov[X_1, X_2] &= Cov[X_1, a + bX_1] \\ &= E[X_1(a + bX_1)] - E[X_1]E[a + bX_1] \\ &= E[aX_1 + bX_1^2] - aE[X_1] - bE[X_1]^2 \\ &= aE[X_1] - aE[X_1] + b(E[X_1^2] - E[X_1]^2) = bV[X_1] \end{aligned}$$

Recall from Chapter 2 that $V[a + bX] = b^2V[X]$. This means that location shifts by an amount a have no effect on the variance, whereas rescaling by b scales the variance by b^2. When $b \neq 0$, the correlation is

$$\frac{bV[X_1]}{\sqrt{V[X_1]}\sqrt{b^2V[X_1]}} = \frac{b}{|b|} = \text{sign}(b) \quad (4.13)^4$$

[4] The sign (b) function returns a value of −1 if b is negative or a value of 1 if b is positive.

When X_1 and X_2 tend to increase together, then the correlation is positive. If X_2 tends to decrease when X_1 increases, then these two random variables have a negative correlation.

The covariance between two random variables scales in an analogous way, so that:

$$\text{Cov}[a + bX_1, c + dX_2] = bd\,\text{Cov}[X_1, X_2]$$

The covariance is defined in terms of the deviation from the mean for each random variable, and so location shifts have no effect. At the same time, the scale of each component contributes multiplicatively to the change in the covariance. Combining these two properties demonstrates that the correlation estimator is scale free:

$$\begin{aligned}\text{Corr}[aX_1, bX_2] &= \frac{ab\,\text{Cov}[X_1, X_2]}{\sqrt{a^2 V[X_1]}\sqrt{b^2 V[X_2]}}\\ &= \frac{ab}{|a||b|}\frac{\text{Cov}[X_1, X_2]}{\sqrt{V[X_1]}\sqrt{V[X_2]}}\\ &= \text{sign}(a)\text{sign}(b)\text{Corr}[X_1, X_2]\end{aligned}$$

The correlation between two random variables is commonly denoted by ρ (or ρ_{12} to specify that this is the correlation between X_1 and X_2).

By rearranging equation (4.12), the covariance can be expressed in terms of the correlation and the standard deviations:

$$\sigma_{12} = \rho_{12}\sigma_1\sigma_2 \quad (4.14)$$

The variance of each component and the covariance between the two components of X are frequently expressed as a 2-by-2 covariance matrix. The covariance matrix of X is:

$$\text{Cov}[X] = \begin{bmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{bmatrix} = \begin{bmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{12} & \sigma_{22}\end{bmatrix} \quad (4.15)$$

Cross-variable versions of the skewness (coskewness) and kurtosis (cokurtosis) are similarly defined by taking powers of the two variables that sum to three or four, respectively.

There are two distinct coskewness measures and three distinct cokurtosis measures. For example, the two coskewness measures are:

$$\frac{E[(X_1 - E[X_1])^2(X_2 - E[X_2])]}{\sigma_1^2\sigma_2} \text{ and } \frac{E[(X_1 - E[X_1])(X_2 - E[X_2])^2]}{\sigma_1\sigma_2^2} \quad (4.16)$$

Like the skewness of a single random variable, the coskewness measures are standardized.

While interpreting these measures is more challenging than the covariance, they both measure whether one random variable (raised to the power 1) takes a clear direction whenever the other return (raised to the power 2) is large in magnitude. For example, it is common for returns on individual stocks to have negative coskew, so that when one return is negative, the other tends to experience a period of high volatility (i.e., its squared value is large).

The Variance of Sums of Random Variables

The covariance plays a key role in the variance of the sum of two or more random variables:

$$V[X_1 + X_2] = V[X_1] + V[X_2] + 2\text{Cov}[X_1, X_2] \quad (4.17)$$

The variance of the sum is the sum of the variances plus twice the covariance. For independent variables, the variance of the sum is simply the sum of the variances. The covariance in the variance of the sum captures the propensity of the two random variables to move together. This result can be generalized to weighted sums:

$$V[aX_1 + bX_2] = a^2V[X_1] + b^2V[X_2] + 2ab\,\text{Cov}[X_1, X_2] \quad (4.18)$$

This result is important in portfolio applications, where X_1 and X_2 might be the returns to two stocks and the constants are the weights in the portfolio.

CORRELATION AND PORTFOLIO DIVERSIFICATION

Correlation plays an important role in determining the benefits of portfolio diversification. The return on a portfolio depends on:

- The distribution of the returns on the assets in the portfolio, and
- The portfolio's weights on these assets.

The portfolio weights measure the share of funds invested in each asset. These weights must sum to 1 by construction, and so in a simple portfolio with two assets, the two portfolio weights are:

- w (i.e., the weight on the first asset), and
- $1 - w$ (i.e., the weight on the second asset).

If the returns on two assets have a covariance matrix of:

$$\begin{bmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{bmatrix}$$

(*Continued*)

(*Continued*)

then the variance of the portfolio's return is:

$$w^2\sigma_1^2 + (1 - w)^2\sigma_2^2 + 2w(1 - w)\sigma_{12}$$

This is an application of the formula for the covariance of weighted sums of random variables. The variance-minimizing weight (i.e., that makes the variance as small as possible) can be shown to be:

$$w^* = \frac{\sigma_2^2 - \sigma_{12}}{\sigma_1^2 - 2\sigma_{12} + \sigma_2^2}$$

This formula can be derived by taking the expression for the variance of the portfolio's return, differentiating it with respect to w, setting the resulting expression to zero, and rearranging it.

Figure 4.2 plots the standard deviation of a portfolio with two assets that have a covariance matrix of:

$$\begin{bmatrix} 20\%^2 & \rho \times 20\% \times 10\% \\ \rho \times 20\% \times 10\% & 10\%^2 \end{bmatrix}$$

In this example, the first asset has equity market-like volatility of 20%, and the second has corporate bond-like volatility of 10%. Meanwhile, the correlation ρ is varied between -1 and 1, and the variance-minimizing portfolio weights w^* and $(1 - w^*)$ are used to calculate the portfolio standard deviation using the square root of the variance.

Note that the portfolio's standard deviation (or variance) can reach its theoretical minimum value of zero when the correlation is either -1 or 1. When the correlation is negative, w^* is positive because the prices of the two assets will tend to move in the opposite direction and so offset each other. Minimizing the variance therefore requires a long position in both assets. On the other hand, when the correlation is positive, w^* is negative as the asset prices tend to move in the same direction and minimizing the variance would require taking offsetting positions (e.g., a long position in asset 1 and a short position in asset 2). It is also asymmetric, because large positive correlations lead to larger portfolio standard deviations than small negative correlations. This happens because the optimal weight is negative for the largest correlations and so the second asset has a weight larger than 1. While it can be optimal to have a weight larger than 1, this extra total exposure limits the benefits to diversification. When the correlation is negative, both assets have weights between 0 and 1, and so this additional source of variance is not present.

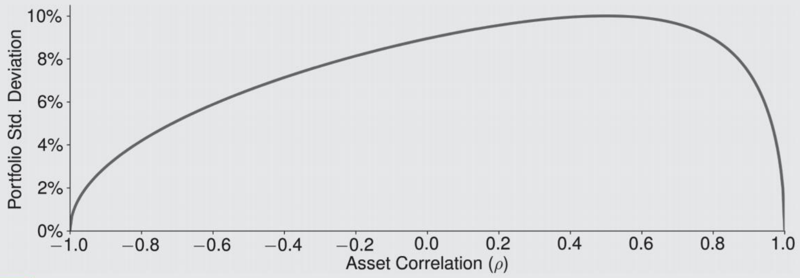


Figure 4.2 **The portfolio standard deviation using the variance-minimizing portfolio weight in a two-asset portfolio, where the first asset has a volatility of 20%, the second has a volatility of 10%, and the correlation between the two is varied between −1 and 1.**

Covariance, Correlation, and Independence

There is an important relationship between correlation and independence. When two random variables are independent, they must have zero correlation because (by definition) each must contain no information about the other. However, if two random variables have zero correlation, they are not necessarily independent. This is because correlation is a measure of *linear* dependence. If two variables have a strong linear relationship (i.e., they produce values that lie close to a straight line), then they have a large correlation. If two random variables have no linear relationship, then their correlation is zero.

However, variables can be non-linearly dependent. A simple example of this occurs when:

$$X_1 \sim N(0,1) \text{ and } X_2 = X_1^2$$

Using the fact that the skewness of a normal is 0, it can be shown that $Cov[X_1, X_2] = 0$. This is because $Cov[X_1, X_2] = Cov[X_1, X_1^2]$, and the latter is simply another way of defining the skewness of X_1. However, these two random variables are clearly not independent. Because the random variable X_2 is a function of X_1, the realization of X_1 determines the realization of X_2. Similarly, a realization of X_2 can be used to predict X_1 up to a multiplicative factor of + 1. Therefore, independence is a stronger property than zero correlation.

Finally, if the correlation between two variables is 0, then the expectation of the product of the two is the product of the expectations. This means that if $Corr[X_1, X_2] = 0$, then $E[X_1X_2] = E[X_1]E[X_2]$. This follows from the definition of covariance because:

$$Cov[X_1, X_2] = E[X_1X_2] - E[X_1]E[X_2],$$

so that when the covariance is zero, the two terms on the right-hand side must be equal

$$E[X_1X_2] = E[X_1]E[X_2].$$

4.3 CONDITIONAL EXPECTATIONS

Many applications examine the expectation of a random variable conditional on the outcome of another random variable. For example, it is common in risk management to model the expected loss on a portfolio given a large negative market return.

A conditional expectation is an expectation when one random variable takes a specific value or falls into a defined range of values. It uses the same expression as any other expectation and is a weighted average where the probabilities are determined by a conditional PMF.

In the stock return/analyst rating example, consider the expected return on the stock given a positive analyst rating. The conditional distribution here [i.e., $f_{X_1|X_2}(x_1|X_2 = 1)$] is:

Stock Return	−5%	0%	5%
Probability	16.6%	16.6%	66.6%

The conditional expectation of the return is then:

$$E[X_1|X_2 = 1] = -5\% \times 16.6\% + 0 \times 16.6\% + 5\% \times 66.6\% \\ = 2.5\%$$

Conditional expectations can be defined over a value or any set of values that X_2 might take. For example, the conditional expectation of X_1 given that the stock is not given a positive rating (i.e., $X_2 \in \{-1,0\}$) depends on the conditional PMF of X_1 given that $X_2 \in \{-1,0\}$. This expectation is computed using the conditional PMF:

$$f_{X_1|X_2}(x_1|X_2 \in \{-1,0\})$$

Conditional expectations can be extended to any moment by replacing all expectation operators with conditional expectation operators. For example, the variance of a random variable X_1 is

$$V[X_1] = E[(X_1 - E[X_1])^2] = E[X_1^2] - E[X_1]^2$$

The conditional variance of X_1 given that X_2 takes some value x_2 is then:

$$V[X_1|X_2 = x_2] = E[(X_1 - E[X_1|X_2 = x_2])^2|X_2 = x_2] \\ = E[X_1^2|X_2 = x_2] - E[X_1|X_2 = x_2]^2.$$

Returning to the example of the stock return and the recommendation, the standard deviation of the stock return (i.e., the square root of the variance) is calculated from the probabilities in the marginal distribution

$$\sigma_1 = \sqrt{V[X_1]}$$
$$= \sqrt{E[X_1^2] - E[X_1]^2}$$
$$= \sqrt{((-0.05)^2(0.35) + 0^2(0.30) + (0.05)^2(0.35)) - (-0.05(0.35) + 0(0.30) + 0.05(0.35))^2}$$
$$= 4.18\%$$

The standard deviation conditional on a positive rating is calculated from the conditional PMF where $X_2 = 1$.

$$\sigma_{X_1|X_2=1} = \sqrt{V[X_1|X_2 = x_2]}$$
$$= \sqrt{E[X_1^2|X_2 = 1] - E[X_1|X_2 = 1]^2}$$
$$= \sqrt{((-0.05)^2(0.166) + 0^2(0.166) + (0.05)^2(0.666)) - (-0.05(0.166) + 0(0.166) + 0.05(0.666))^2}$$
$$= 3.81\%$$

Conditional Independence

As a rule, random variables are dependent in finance and risk management. This dependence arises from many sources, including shifts in investor risk aversion, cross-asset or cross-border spillovers, and crowded portfolio strategies. However, conditioning is a useful tool to remove the dependence between variables.

Returning to the stock return and recommendation example, suppose that an investor wants to identify firms that have a conservative management style. This style variable can be labeled Z.

Now suppose that 12.5% of the companies are in the conservative category and that the probability of the outcomes of stock

returns and recommendations based on management style are as in the table below:

	Style (Z)		Conservative			Other		
			Stock Return (X_1)			Stock Return (X_1)		
			−5%	0%	5%	−5%	0%	5%
Analyst (X_2)	Negative	−1	1%	4%	0%	19%	6%	0%
	Neutral	0	1%	4%	0%	9%	11%	15%
	Positive	1	0.5%	2%	0%	4.5%	3%	20%

Note that the overall probabilities are consistent with the original bivariate matrix. For example, there is still a total 20% chance of a negative rating when the stock return is −5%.

Now consider the bivariate distribution conditional on a company being in this conservative category. As before, each individual probability is rescaled by a constant factor so that the total conditional probability sums to 1. In other words, each value is divided by the probability that the management is conservative (i.e., 12.5%), which is also the sum of the values in the panel.

			Conservative			
			Stock Return (X_1)			
			−5%	0%	5%	$f_{X_2}(x_2)$
Analyst (X_2)	Negative	−1	8%	32%	0%	40%
	Neutral	0	8%	32%	0%	40%
	Positive	1	4%	16%	0%	20%
		$f_{X_1}(x_1)$	20%	80%	0%	

This shows that this bivariate distribution, which conditions on the company having a conservative management style, is equal to the product of its marginal distributions. To see this, check that the cells of the matrix are equal to the product of their corresponding marginals [e.g., 8% = 40%(20%)]. This distribution is therefore conditionally independent even though the original distribution is not.

Mathematically, this conditional independence can be represented

$$f_{X1|X2}(X_1|X_2 = 12.5\%) = f_{X1}(X_1)f_{X2}(X_2)$$

4.4 CONTINUOUS RANDOM VARIABLES

As is the case with univariate random variables, moving from discrete multivariate random variables to those that are continuous changes little. The most substantial change is the replacement of the PMF with a PDF, and the corresponding switch from sums to integrals when computing expectations. All definitions and high-level results (e.g., the effect of linear transformations on the covariance of random variables) are identical.

The PDF of a bivariate continuous random variable returns the probability associated with a single point. This value is not technically a probability because, like a univariate random variable, it is only possible to assign a positive probability to intervals and not to a single outcome. The PDF has the same form as a PMF and is written as $f_{X_1,X_2}(x_1, x_2)$. The probability in a rectangular region is defined as:

$$\Pr(l_1 < X_1 < u_1 \cap l_2 < X_2 < u_2) = \int_{l_1}^{u_1}\int_{l_2}^{u_2} f_{X_1,X_2}(x_1, x_2)dx_2\, dx_1 \tag{4.19}$$

This expression computes the area under the PDF function in the region defined by l_1, u_1, l_2, and u_2.[5] A PDF function is always non-negative and must integrate to one across the support of the two components so that the area under the PDF is one.

The cumulative distribution function is the area of a rectangular region under the PDF where the lower bounds are $-\infty$ and the upper bounds are the arguments in the CDF:

$$F_{X_1,X_2}(x_1, x_2) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2} f(t_1, t_2)dt_2\, dt_1 \tag{4.20}$$

If the PDF is not defined for all values, then the lower bounds can be adjusted to be the smallest values where the random variable has support.

Most continuous multivariate random variables have complicated PDFs. The simplest is the bivariate uniform where the two components are independent. The PDF of this random variable is constant:

$$f_{X_1,X_2}(x_1, x_2) = 1 \tag{4.21}$$

and the support of the density is the unit square (i.e., a square whose sides have lengths of 1). The CDF of the bivariate independent uniform is:

$$F_{X_1,X_2}(x_1, x_2) = x_1x_2 \tag{4.22}$$

This PDF and CDF correspond to a uniform random variable that has independent components. This is not always the case, however, and there are many examples of uniform random variables that are dependent. Dependent uniforms and the special role they play in understanding the dependence in any multivariate random variable are examined in later chapters.

[5] The probability that a continuous random variable takes a single point value is 0, and so the probability over a closed interval is the same as the probability over the open interval with the same bounds, so that $\Pr(l_1 < X_1 < u_1 \cap l_2 < X_2 < u_2) = \Pr(l_1 \le X_1 \le u_1 \cap l_2 \le X_2 \le u_2)$.

Marginal and Conditional Distributions

For a continuous bivariate random variable, the marginal PDF integrates one component out of the joint PDF.

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1, x_2)dx_2 \quad (4.23)$$

This definition is identical to that of the marginal distribution of a discrete random variable, aside from the replacement of the sum with an integral. The effect is the also same, and the marginal PDF of one component is just a continuous univariate random variable.

The conditional PDF is analogously defined, and the conditional PDF of X_1 given X_2 is the ratio of the joint PDF to the PDF of X_2:

$$f_{X_1|X_2}(x_1|X_2 = x_2) = \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)} \quad (4.24)$$

This conditional PDF is the distribution of X_1 if X_2 is known to be equal to x_2. The conditional PDF is a density and has all the properties of a univariate PDF.

The conditional distribution can also be defined when X_2 is in an interval. Conditional distributions of this form are important in risk management when X_1 and X_2 are asset returns. In this case, the distribution of interest is that of X_1 given that X_2 has suffered a large loss. This structure is used in later chapters when examining concepts such as expected shortfall. For example, consider the distribution of the return on a hedge fund (X_1) given that the monthly return on the S&P 500 (X_2) is in the bottom 5% of its distribution. Historical data show that the S&P 500 return is above −6.19% in 95% of months, and thus the case of interest is $X_2 \in (-\infty, -6.19\%)$. The conditional PDF of the hedge fund's return is

$$f_{X_1|X_2}(x_1|X_2 \in (-\infty, -6.19\%)) = \frac{\int_{-\infty}^{-0.0619} f_{X_1,X_2}(x_1, x_2)dx_2}{\int_{-\infty}^{-0.0619} f_{X_2}(x_2)dx_2}$$

While this expression might look strange, it is identical to the definition of the PMF when one of the variables is in a set of values. This conditional PDF is a weighted average of the PDF of X_1 when $X_2 < -6.19\%$, with the weights being the probabilities associated with the values of X_2 in this range.

4.5 INDEPENDENT, IDENTICALLY DISTRIBUTED RANDOM VARIABLES

Independent, identically distributed (iid) random variables are those that are generated by a single univariate distribution. They are commonly employed as fundamental shocks in models, especially in time series. For example, it is common to write:

$$X_i \overset{iid}{\sim} N(\mu, \sigma^2),\ i = 1,2, \ldots ,n,$$

as a generator of an n-component multivariate normal random variable where each component is independent of all other components and is distributed $N(\mu, \sigma^2)$.

Manipulating iid random variables is particularly simple because the variables are independent and have the same moments. For example, the expected value of a sum of n iid random variables is:

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} \mu = n\mu, \quad (4.25)$$

where $\mu = E[X_i]$ is the common mean of the random variables.

This result only depends on the fact that the variables are identical because the expectation of a sum is always the sum of the expectations (due to the linearity of the expectation operator).

The variance of a sum of n iid random variables is

$$\begin{aligned} V\left[\sum_{i=1}^{n} X_i\right] &= \sum_{i=1}^{n} V[X_i] + 2\sum_{j=1}^{n}\sum_{k=j+1}^{n} Cov[X_j, X_k] \\ &= \sum_{i=1}^{n} \sigma^2 + 2\sum_{j=1}^{n}\sum_{k=j+1}^{n} 0 \\ &= n\sigma^2 \end{aligned} \quad (4.26)$$

The variance of a sum of iid random variables is just n multiplied by the variance of one variable since each variable has the same variance σ^2 and the covariances are all zero.

To see this, note that the first line shows the variance of the sum to be the sum of the variances plus *all* pairwise covariances multiplied by two. There are $n(n - 1)/2$ distinct pairs, and the double sum is a simple method to account for the distinct combination of indices. Furthermore, the independence assumption ensures that the covariances are all zero, which leads to the simplification of the sum. If the data are not independent, then the covariance between different components would not be zero and would therefore appear in the variance of the sum.

The assumption that the distribution of X_i is identical is also important because it ensures that $V[X_i] = \sigma^2$ for all i. Without this assumption, the variance of the sum is still the sum of the variances, although each variance may take a different value.

It is important to distinguish between the variance of the sum of multiple random variables and the variance of a multiple of a single random variable. In particular, if X_1 and X_2 are iid with variance σ^2, then:

$$V[X_1 + X_2] = 2\sigma^2$$

is different from

$$V[2X_1] = 4\sigma^2$$

or

$$V[2X_2] = 4\sigma^2$$

This property plays an important role when estimating unknown parameters. The variance of the sum of iid random variables grows linearly. This means that when the sum of n random variables is divided by n to form an average, the variance of the average reduces as n grows. This property is explored in detail in the next chapter.

4.6 SUMMARY

Multivariate random variables are natural extensions of univariate random variables. They are defined using PMFs (for discrete variables) or PDFs (for continuous variables), which describe the joint probability of outcome combinations. Expected values are computed using the same method, where the outcomes are weighted by their associated probabilities. The same moments used for univariate random variables (i.e., the mean and the variance) are also useful for describing bivariate random variables. However, these moments are not sufficient to fully describe a bivariate distribution, and so it is necessary to describe the dependence between the components. The most common measure of dependence is the covariance, which measures the common dispersion between two variables. In practice, it is more useful to report the correlation, which is a scale- and unit-free transformation of the covariance.

Bivariate distributions can be transformed into either marginal or conditional distributions. A marginal distribution summarizes the information about a single variable and is simply a univariate distribution. A conditional distribution describes the probability of one random variable conditional on an outcome or a range of outcomes of another. The conditional distribution plays a key role in finance and risk management, where it is used to measure the return on an asset conditional on a significant loss in another asset. The conditional distribution can also be used to produce a conditional expectation and conditional moments, which provide summaries of the conditional distribution of a random variable.

Moving from discrete to continuous random variables makes little difference aside from some small technical changes. The concepts are identical, and all properties of moments continue to hold.

The following questions are intended to help candidates understand the material. They are _not_ actual FRM exam questions.

QUESTIONS

Short Concept Questions

4.1 How are the marginal distributions related when two random variables are independent?

4.2 When are conditional distributions equal to marginal distributions?

4.3 If X_1 and X_2 are independent, what is the correlation between X_1 and X_2?

4.4 If X_1 and X_2 are uncorrelated ($\text{Corr}[X_1, X_2] = 0$), are they independent?

4.5 What is the effect on the covariance between X_1 and X_2 of rescaling X_1 by w and X_2 by $(1 - w)$, where w is between 0 and 1?

4.6 Under what conditions are conditional expectations and marginal expectations equal?

4.7 If X_1 and X_2 are independent continuous random variables with PDFs $f_{X_1}(x_1)$ and $f_{X_2}(x_2)$, respectively, what is the joint PDF of the two random variables?

4.8 If X_1 and X_2 both have univariate normal distributions, is the joint distribution of X_1 and X_2 a bivariate normal?

4.9 In the previous question, what if X_1 and X_2 are independent?

4.10 Are sums of iid normal random variables normally distributed?

Practice Questions

4.11 Suppose that the annual profit of two firms, one an incumbent (Big Firm, X_1) and the other a startup (Small Firm, X_2), can be described with the following probability matrix:

		Small Firm (X_2)			
		−USD 1M	USD 0	USD 2M	USD 4M
Big Firm (X_1)	−USD 50M	1.97%	3.90%	0.8%	0.1%
	USD 0	3.93%	23.5%	12.6%	2.99%
	USD 10M	0.8%	12.7%	14.2%	6.68%
	USD 100M	0%	3.09%	6.58%	6.16%

a. What are the marginal distributions of the profits of each firm?

b. Are the returns on the two firms independent?

c. What is the conditional distribution of Small Firm's profits if Big Firm has a $100M profit year?

4.12 Using the probability matrix for Big Firm and Small Firm, what are the covariances and correlations between the profits of these two firms?

4.13 If an investor owned 20% of Big Firm and 80% of Small Firm, what are the expected profits of the investor and the standard deviation of the investor's profits?

4.14 In the Big Firm–Small Firm example, what are the conditional expected profit and conditional standard deviation of the profit of Big Firm when Small Firm either has no profit or loses money ($X_2 \leq 0$)?

4.15 The expected return distributions for the S&P 500 and Nikkei for the next year are given as:

S&P 500	Return	−10%	0%	+10%
	Probability	25%	50%	25%
Nikkei	Return	−5%	0%	8%
	Probability	20%	60%	20%

a. What is the joint distribution matrix if the two return series are unrelated?

b. If the actual matrix looks as follows:

		S&P 500		
		−10%	0%	+10%
Nikkei	−5%	15%		1%
	0%	5%	40%	
	8%		6%	9%

Fill in the missing cells to match the original given marginal distributions.

c. For the matrix in b, what is the conditional distribution of the Nikkei given a 10% return on the S&P?

The following questions are intended to help candidates understand the material. They are *not* actual FRM exam questions.

ANSWERS

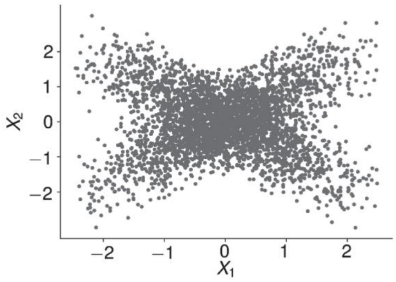
Short Concept Questions

4.1 The joint PMF is the product of the two marginals. This is a key property of independence.

4.2 When the random variables are independent. If the conditional distribution of X_1 given X_2 equals the marginal, then X_2 has no information about the probability of different values in X_1.

4.3 Independent random variables always have a correlation of 0. This statement is only technically true if the variance of both is well defined so that the correlation is also well defined.

4.4 Not necessarily. There are many ways that two random variables can be dependent but not uncorrelated. For example, if there is no linear relationship between the variables but they both tend to be large in magnitude (of either sign) at the same time. In this case, there may be tail dependence but not linear dependence (correlation). The image below plots realizations from a dependent bivariate random variable with 0 correlation. The dependence in this distribution arises because if X_1 is close to 0 then X_2 is also close to 0.

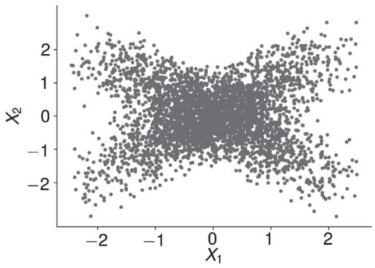


4.5 Using the properties of rescaled covariance, $Cov[wX_1, (1 - w)X_2] = w(1 - w)Cov[X_1, X_2]$.

4.6 If the two components X_1 and X_2 are independent, then the conditional expectation of any function of X_1 given $X_2 = x_2$ is always the same as the marginal expectation of the same function. That is, $E[g(X_1)|X_2 = x_2] = E[g(X_1)]$ for any value x_2.

4.7 The joint PDF is $f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1)f_{X_2}(x_2)$ because when independent the joint is the product of the marginals.

4.8 Not necessarily. It is possible that the joint is not a bivariate normal if the dependence structure between X_1 and X_2 is different from what is possible with a normal. The distribution that generated the data points plotted below has normal marginals, but this pattern of dependence is not possible in a normal that always appears elliptical.



4.9 The joint distribution would be bivariate normal if these are independent, because this is a bivariate normal with 0 correlation.

4.10 Yes because iid normal random variables are jointly normally distributed with 0 correlation, the sum of iid normal random variables is also normal.

Solved Problems

4.11 **a.** The marginal distributions are computed by summing across rows for Big Firm and down columns for Small Firm. They are:

Big Firm		Small Firm	
−USD 50M	6.77%	−USD 1M	6.70%
USD 0	43.02%	USD 0	43.19%
USD 10M	34.38%	USD 2M	34.18%
USD 100M	15.83%	USD 4M	15.93%

So, for example, the first entry for big firms is 1.97% + 3.90% + 0.8% + 0.1% = 6.77% and the first entry for small firms is 1.97% + 3.93% + 0.8% + 0% = 6.70%.

b. They are independent if the joint is the product of the marginals for all entries in the probability matrix. Looking at the upper left cell, 6.77% × 6.70% is not equal to 1.97%, and so they are not.

c. The conditional distribution is just the row corresponding to USD 100M normalized to sum to 1. The conditional is different from the marginal, which is another demonstration that the profits are not independent.

−USD 1M	USD 0	USD 2M	USD 4M
0%	19.5%	41.6%	38.9%

For example, the second entry in this table is calculated as

100% × [3.09%/(3.09% + 6.58% + 6.16%)] = 19.5%

4.12 First, the two means and variances can be computed from the marginal distributions in the earlier problem. Then the covariances can be computed using the alternative form, which is $E[X_1X_2] - E[X_1]E[X_2]$. The means can be computed using $E[X_j] = \Sigma x_j \Pr(X_j = x_j)$ for $j = 1,2$. The variance can be computed using $E[X_j^2] - (E[X_j])^2$, which requires computing $E[X_j^2]$ using $\Sigma x_j^2 \Pr(X_j = x_j)$.

For Big Firm, these values are $E[X_1] = 6.77\% \times (-50) + 4.302\% \times (0) + 34.38\% \times (10) + 15.83\% \times (100) = \$15.88M$, $E[X_1^2] = 1786.63$ and $V[X_1] = 1534.36$.

For Small Firm, these values are $E[X_2] = 6.77\% \times (-50)^2 + 4.302\% \times (0)^2 + 34.38\% \times (10)^2 + 15.83\% \times (100)^2 = \$1.25M$, $E[X_2^2] = 3.98$ and $V[X_2] = 2.41$.

The expected value of the cross product is

$$E[X_1X_2] = \Sigma\Sigma x_1x_2 \Pr(X_1 = x_1, X_2 = x_2) = 43.22.$$

This is obtained by multiplying each of the 16 combinations of payoffs to the two firms by their respective probabilities in the matrix and then summing them

$\$-50M \times \$-1M \times 1.97\% + \$-50M \times \$0M \times 3.90\% + \ldots + \$100M \times \$4M \times 6.16\%$ and finally dividing buy 100% at the end. The covariance is $E[X_1 X_2] - E[X_1] E[X_2] = 43.22 - 15.88 \times 1.25 = 23.37$ and the correlation is $23.37/(\sqrt{2.41 \times 1534.36} = 0.384$

4.13 Here we can compute these from the components in the previous step. The expected profits are then $E[P] = 20\% \times E[X_1] + 80\% \times E[X_2] = \$4.18M$. The variance of the portfolio is then

$$(20\%)^2 Var[X_1] + (80\%)^2 Var[X_2] + 2(20\%)(80\%)Cov[X_1, X_2]$$

This value is $0.04 \times 1534.36 + 0.64 \times 2.41 + 2 \times 0.16 \times 23.37 = 70.39$, and so the standard deviation is USD $8.39M$.

4.14 We need to compute the conditional distribution given $X_2 \le 0$. The relevant rows of the probability matrix are

X_1/X_2	−USD 1M	USD 0
−USD 50M	1.97%	3.90%
USD 0	3.93%	23.5%
USD 10M	0.8%	12.7%
USD 100M	0%	3.09%

The conditional distribution can be constructed by summing across rows and then normalizing to sum to unity. The non-normalized sum and the normalized version are

$X_1 \mid X_2 \le 0$	Non-Normalized	Normalized
−USD 50M	5.87%	11.77%
USD 0	27.43%	54.98%
USD 10M	13.5%	27.06%
USD 100M	3.09%	6.19%

Finally, the conditional expectation is $E[X_1 \mid X_2 \le 0] = \Sigma x_1 \Pr(X_1 = x_1 \mid X_2 \le 0) =$ USD $3.01M$. The conditional expectation squared is $E[X_1^2 \mid X_2 \le 0] = 940.31$, and so the conditional variance is $V[X_1] = E[X_1^2] - E[X_1]^2 = 940.31 - 3.01^2 = 931.25$ and the conditional standard deviation is USD $30.52M$.

4.15 a. If the two variables are unrelated, then the joint distribution is just given by the products

$$f_{X_1, X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$$

For example, the probability of seeing −10% on the S&P and −5% on the Nikkei is 25%*20% = 5%:

		S&P 500		
		−10%	**0%**	**+10%**
Nikkei	**−5%**	**5%**	**10%**	**5%**
	0%	**15%**	**30%**	**15%**
	8%	**5%**	**10%**	**5%**

b.

		S&P 500		
		−10%	**0%**	**+10%**
Nikkei	**−5%**	**15%**		**1%**
	0%	**5%**	**40%**	
	8%		**6%**	**9%**

The sum of the rows needs to match the S&P 500 marginal distribution:

S&P 500	Return	**−10%**	**0%**	**+10%**
	Probability	25%	50%	25%

In other words, the x in this column:

	−10%
−5%	**15%**
0%	**5%**
8%	**x%**

needs to be such that the sum of $15 + 5 + x = 25$. So, $x = 5\%$.

Continuing yields:

		S&P 500		
		−10%	**0%**	**+10%**
Nikkei	**−5%**	**15%**	**4%**	**1%**
	0%	**5%**	**40%**	**15%**
	8%	**5%**	**6%**	**9%**

The same process works going across the columns.

c. Working with the last column:

		S&P 500
		+10%
Nikkei	**−5%**	**1%**
	0%	**15%**
	8%	**9%**
		25%

Dividing each entry by the total mass (25% in this case) gives the conditional distribution:

		S&P 500	Conditional Distribution
		+10%	
Nikkei	**−5%**	**1%**	**4%**
	0%	**15%**	**60%**
	8%	**9%**	**36%**
		25%	**100%**

Sample Moments

5

Learning Objectives

After completing this reading you should be able to:

- Estimate the mean, variance, and standard deviation using sample data.
- Explain the difference between a population moment and a sample moment.
- Distinguish between an estimator and an estimate.
- Describe the bias of an estimator and explain what the bias measures.
- Explain what is meant by the statement that the mean estimator is BLUE.
- Describe the consistency of an estimator and explain the usefulness of this concept.
- Explain how the Law of Large Numbers (LLN) and Central Limit Theorem (CLT) apply to the sample mean.
- Estimate and interpret the skewness and kurtosis of a random variable.
- Use sample data to estimate quantiles, including the median.
- Estimate the mean of two variables and apply the CLT.
- Estimate the covariance and correlation between two random variables.
- Explain how coskewness and cokurtosis are related to skewness and kurtosis.

This chapter describes how sample moments are used to estimate unknown population moments. In particular, this chapter pays special attention to the estimation of the mean. This is because when data are generated from independent and identically distributed (iid) random variables, the mean estimator has several desirable properties.

- It is (on average) equal to the population mean.
- As the number of observations grows, the sample mean becomes arbitrarily close to the population mean.
- The distribution of the sample mean can be approximated using a standard normal distribution.

This final property is widely used to test hypotheses about population parameters using observed data.

Data can also be used to estimate higher-order moments such as variance, skewness, and kurtosis. The first four (standardized) moments (i.e., mean, variance, skewness, and kurtosis) are widely used in finance and risk management to describe the key features of data sets.

Quantiles provide an alternative method to describe the distribution of a data set. Quantile measures are particularly useful in applications to financial data because they are robust to extreme outliers. Finally, this chapter shows how univariate sample moments can extend to more than one data series.

5.1 THE FIRST TWO MOMENTS

Estimating the Mean

When working with random variables, we are interested in the value of population parameters such as the mean (μ) or the variance (σ^2). However, these values are not observable, and so data are used to estimate these values. The population mean is estimated using the sample mean (i.e., average) of the data. The mean estimator is defined as:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{5.1}$$

A mean estimate $\hat{\mu}$ is obtained from Equation 5.1 by replacing the random variables (i.e., X_i) with their observed values (i.e., x_i). The hat-notation ($\hat{\ }$) is used to distinguish an estimator or a sample estimate (in this case $\hat{\mu}$) from the unknown population parameter (in this case μ) of interest. Note that $\overline{X}$ is another common symbol for the sample mean. In this chapter, the random variables X_i are assumed to be independent and identically distributed, so that $E[X_i] = \mu$ and $V[X_i] = \sigma^2$ for all i.

The mean estimator is a function that transforms the data into an estimate of the population mean. More generally, an estimator is a mathematical procedure that calculates an estimate based on an observed data set. In contrast, an estimate is the value produced by an application of the estimator to data.[1]

The mean estimator is a function of random variables, and so it is also a random variable. Its properties can be examined using the tools developed in the previous chapters.

The expectation of the mean is

$$E[\hat{\mu}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu \tag{5.2}$$

The expected value of the mean estimator is the same as the population mean.[2] While the primary focus is on the case where X_i are iid, this result shows that the expectation of the mean estimator is μ whenever the mean is constant (i.e., $E[X_i] = \mu$ for all i). This property is useful in applications to time-series that are not always iid but have a constant mean.

The bias of an estimator is defined as:

$$\text{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta \tag{5.3}$$

where θ is the true (or population) value of the parameter that we are estimating (e.g., μ or σ^2). It measures the difference between the expected value of the estimator and the population value estimated. Applying this definition to the mean:

$$\text{Bias}(\hat{\mu}) = E[\hat{\mu}] - \mu = \mu - \mu = 0$$

Because the mean estimator's bias is zero, it is unbiased.

The variance of the mean estimator can also be computed using the standard properties of random variables.[3] Recall that the general formula for the variance of a sum is the sum of the variances plus any covariances between the random variables:

$$V[\hat{\mu}] = V\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}V\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\left\{\sum_{i=1}^{n} V[X_i] + \text{Covariances}\right\} \tag{5.4}$$

Since X_i are iid, these variables are uncorrelated and so have 0 covariance. Thus:

$$V\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} V[X_i] = \frac{1}{n^2}(n\sigma^2) = \sigma^2/n \tag{5.5}$$

[1] This definition only covers point estimators, that is, an estimator whose value is a single point (a scalar or finite-dimensional vector). This is the most widely used class of estimators.

[2] This result only depends on the linearity of the expectation operator, so that the expectation of the sum is the sum of the expectation.

[3] The variance of the sample mean estimator is distinct from variance of the underlying data (i.e., $X_1, X_2, \ldots X_n$). The variance of the sample mean depends on the sample size n, whereas the variance of the data does not. These two are commonly confused when first encountered because the variance of the sample mean depends on the variance of the data (i.e., σ^2). See the box, *Standard Errors and Standard Deviation*, for more discussion about the distinction between the variance of an estimator and the variance of the data used in the estimator.

The variance of the mean estimator depends on two values: the variance of the data (i.e., σ^2) and the number of observations (i.e., n). The variance in the data is noise that obscures the mean. The more variable the data, the harder it is to estimate the mean of that data. The variance of the mean estimator also decreases as the number of observations increases, and so larger samples produce estimates of the mean that tend to be closer to the population mean. This occurs because there is more information in the sample when the number of data points is larger, increasing the accuracy of the estimated mean.

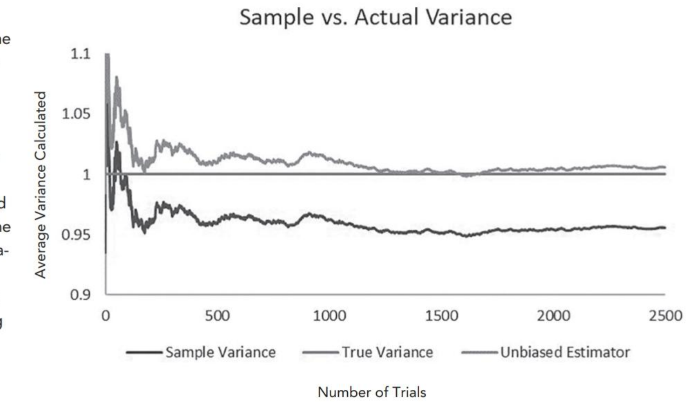


Figure 5.1 **The figure shows the average variance, calculated using the sample variance and the unbiased estimator, for an increasing number of samples of size 20. The values of the biased and unbiased estimators converge toward 0.95 and 1, respectively, as the number of samples increases.**

Estimating the Variance and Standard Deviation

Recall that the variance of a random variable is defined as:

$$\sigma^2 = V[X] = E[(X - E[X])^2]$$

Given a set of n iid random variables X_i, the sample estimator of the variance is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2 \quad (5.6)$$

A sample variance estimate (i.e., the value of $\hat{\sigma}^2$) is obtained from Equation 5.6 by replacing the X_i with the observed values x_i.

Note that the sample variance estimator and the definition of the population variance have a strong resemblance. Replacing the expectation operator E[.] with the averaging operator $n^{-1}\sum_{i=1}^{n}[.]$ transforms the expression for the population moment into the estimator moment. The sample average is known as the sample analog to the expectation operator. It is a powerful tool that is used throughout statistics, econometrics, and this chapter to transform population moments into estimators.[4]

Unlike the mean, the sample variance is a biased estimator. It can be shown that:

$$E[\hat{\sigma}^2] = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2 \quad (5.7)$$

The sample variance is therefore biased with:

$$\begin{aligned}\text{Bias}(\hat{\sigma}^2) &= E[\hat{\sigma}^2] - \sigma^2 \\ &= \frac{n-1}{n}\sigma^2 - \sigma^2 \\ &= \sigma^2/n\end{aligned}$$

Note that this bias is small when n is large.[5] The bias arises because the sample variance depends on the estimator of the mean. Estimation of the mean consumes a degree of freedom, and the sample mean tends to resemble the sample of data a little too well (especially for smaller sample sizes). This slight "overfitting" of the observed data produces a slight underestimation of the population variance.

Because the bias is known, an unbiased estimator for the variance can be constructed as:

$$s^2 = \frac{n}{n-1}\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2 \quad (5.8)$$

Note that this procedure divides by $n - 1$, rather than by n, and the resulting estimator is unbiased since

$$E[s^2] = \sigma^2$$

To see this principle in action, suppose that multiple samples are drawn from an N(0,1) distribution, each sample has a size of 20, and a sample variance $\hat{\sigma}^2$ is calculated for each sample. The results are as shown in Figure 5.1.

As expected, the value converges to 0.95 as the number of samples of 20 increases because:

$$E[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2 = \frac{19}{20}(1) = 0.95$$

Also shown is the unbiased variance estimator s^2, which is simply the biased estimator multiplied by $\frac{n}{n-1}$.

[4] Recall that the population variance can be equivalently defined as $E[X^2] - E[X]^2$. The sample analog can be applied here, and an alternative estimator of the sample variance is $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}X_i^2 - (n^{-1}\sum_{j=1}^{n}X_j)^2$. This form is numerically identical to Equation 5.6.

[5] This is an example of a finite sample bias, and $\hat{\sigma}^2$ is asymptotically unbiased because $\lim_{n\to\infty}\text{Bias}(\hat{\sigma}^2) = 0$.

STANDARD ERRORS VERSUS STANDARD DEVIATIONS

The variance of the sample mean depends on the variance of the data. The standard deviation of the mean estimator is

$$\sqrt{\sigma^2/n} = \sigma/\sqrt{n}$$

and thus, it depends on the standard deviation of the data.

The standard deviation of the mean (or any other estimator) is known as a standard error. Standard errors and standard deviations are similar, although these terms are not used interchangeably. Standard deviation refers to the uncertainty in a random variable or data set, whereas standard error is used to refer to the uncertainty of an estimator. This distinction is important, because the standard error of an estimator declines as the sample size increases. Meanwhile, standard deviation does not change with the sample size.

Standard errors are important in hypothesis testing and when performing Monte Carlo simulations. In both of these applications, the standard error provides an indication of the accuracy of the estimate.

The difference in the bias of these two estimators might suggest that s^2 is a better estimator of the population variance than $\hat{\sigma}^2$. However, this is not necessarily the case because $\hat{\sigma}^2$ has a smaller variance than s^2 (i.e., $V[\hat{\sigma}^2] < V[s^2]$). Financial statistics typically involve large datasets, and so the choice between $\hat{\sigma}^2$ or s^2 makes little difference in practice. The convention is to prefer $\hat{\sigma}^2$ when the sample size is moderately large (i.e., $n \geq 30$).

The sample standard deviation is estimated using the square root of the sample variance (i.e., $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ or $s = \sqrt{s^2}$). The square root is a nonlinear function and so both estimators of the standard deviation are biased. However, this bias diminishes as n becomes large and is typically small in large financial data sets.

Presenting the Mean and Standard Deviation

Means and standard deviations are the most widely reported statistics. Their popularity is due to several factors.

- The mean and standard deviation are often sufficient to describe the data (e.g., if the data are normally distributed or are generated by some other one- or two-parameter distributions).
- These two statistics provide guidance about the likely range of values that can be observed.
- The mean and standard deviation are in the same units as the data, and so can be easily compared. For example, if the data are percentage returns of a financial asset, the mean and standard deviation are also measured in percentage returns. This is not true for other statistics, such as the variance.

SCALING OF THE MEAN AND STANDARD DEVIATION

The most familiar and straightforward way to calculate returns is to take the difference between the current and previous price and divide it by the previous price. But returns can also be measured using the difference between log prices (i.e., the natural logarithm of the returns). Log returns are convenient, as the two-period return is just the sum of two consecutive returns:

$$\ln P_2 - \ln P_0 = (\ln P_2 - \ln P_1) + (\ln P_1 - \ln P_0),$$
$$= R_2 + R_1$$

where P_i is the price of the asset in period i and the return R_i is defined using consecutive prices. Note that log-returns can be validly summed over time, whereas simple returns cannot. The convention followed here is that the first price occurs at time 0 so that the first return, which is measured using the price at the end of period one, is R_1.

The return over n periods is then:

$$R_1 + R_2 + \cdots + R_n = \sum_{i=1}^{n} R_i$$

If returns are iid with mean $E[R_i] = \mu$ and variance $V[R_i] = \sigma^2$, then the mean and variance of the n-period return are $n\mu$ and $n\sigma^2$, respectively. Note that the mean follows directly from the expectation of sums of random variables, because the expectation of the sum is the sum of the expectation.

The variance relies crucially on the iid property, so that the covariance between the returns on a given asset over time is 0. In practice, financial returns are not iid but can be uncorrelated, which is sufficient for this relationship to hold. Finally, standard deviation is the square root of the variance, and so the n-period standard deviation scales with $\sqrt{n}$.

One challenge when using asset price data is the choice of sampling frequency. Most assets are priced at least once per day, and many assets have prices that are continuously available throughout the trading day (e.g., equities, some sovereign bonds, and futures). Other return series, such as hedge fund returns, are only available at lower frequencies (e.g., once per month or even once per quarter or year). This can create challenges when describing the mean or standard deviation of financial returns. For example, sampling over one day could give an asset's average return (i.e., $\hat{\mu}_{Daily}$) as 0.1%, whereas sampling over one week (i.e., $\hat{\mu}_{Weekly}$) could indicate an average return of 0.485%.

In practice, it is preferred to report the annualized mean and standard deviation, regardless of the sampling frequency. Annualized sample means are scaled by a constant that measures the number of sample periods in a year. For example, a monthly mean is multiplied by 12 to produce an annualized mean. Similarly, a weekly mean is multiplied by 52 to produce an annualized version, and

a daily mean is multiplied by the number of trading days per year (e.g., 252 or 260). Note that the scaling convention for daily statistics varies by country (and possibly by the year) because the number of trading days differs across markets.

Returning to the previous example with the samples from one day and one week:

$$\begin{aligned}\hat{\mu}_{Annual} &= 252\hat{\mu}_{Daily} = 52\hat{\mu}_{Weekly}\\ &= (252)(0.001) = (52)(0.00485)\\ &= 2.52\%\end{aligned}$$

Meanwhile, standard deviations use the square root of the same scale factors because standard deviations scale with the square root of the time interval. For example, the standard deviation computed from daily data is multiplied by $\sqrt{252}$ to produce an annualized standard deviation.

Estimating the Mean and Variance Using Data

As an example, consider a set of four data series extracted from the Federal Reserve Economic Data (FRED) database of the Federal Reserve Bank of St. Louis:[6]

1. The ICE BoAML US Corp Master Total Return Index Value, which measures the return on a diversified portfolio of corporate bonds;
2. The Russell 2000 Index, a small-cap equity index that measures the average performance of 2,000 firms;
3. The Gold Fixing Price in the London bullion market, the benchmark price for gold; and
4. The spot price of West Texas Intermediate Crude Oil, the leading price series for US crude oil.

All data in this set are from the period between February 1, 1989 and December 28, 2018. These prices are sampled daily, weekly (on Thursdays), and at the end of the month.[7] Returns are computed using log returns, defined as the difference between consecutive log prices (i.e., $\ln P_t - \ln P_{t-1}$).

Table 5.1 reports the means and standard deviations for these four series. Note how the values change across the three sampling intervals. The bottom rows of each panel contain the annualized versions of each statistic, which largely remove the effect of the sampling frequency. The scale factors applied are 245 (for daily returns), 52 (for weekly returns), and 12 (for monthly returns). The annual versions of the means are quite close to one another, as are the annualized standard deviations.

5.2 HIGHER MOMENTS

Beyond the mean and standard deviation, two higher order moments are also commonly measured in financial data: skewness and kurtosis.

Recall that the skewness is a standardized version of the third central moment, and so it is unit- and scale-free. The population value of the skewness is defined as:

$$\text{Skewness}(X) = \frac{E[(X - E[X])^3]}{E[(X - E[X])^2]^{3/2}} = \frac{\mu_3}{\sigma^3}, \quad (5.9)$$

where μ_3 is the third central moment and σ is the standard deviation.

The third moment raises deviations from the mean to the third power. This has the effect of amplifying larger shocks relative to smaller ones (compared to the variance, which only squares deviations). Because the third power also preserves the sign of the deviations, an asymmetric random variable will not have a third moment equal to 0. When large deviations tend to come from the left tail, the distribution is negatively skewed. If they are more likely to come from the right tail, then the distribution is positively skewed.

Kurtosis is similarly defined using the fourth moment standardized by the squared variance. It is also unit- and scale-free:

$$\text{Kurtosis}(X) = \kappa = \frac{E[(X - E[X])^4]}{E[(X - E[X])^2]^2} = \frac{\mu_4}{\sigma^4} \quad (5.10)$$

The kurtosis uses a higher power than the variance, and so it is more sensitive to large observations (i.e., outliers). It discards sign information (because a negative deviation from the mean becomes positive when raised to the fourth power), and so it measures the relative likelihood of seeing an observation from the tail of the distribution.

The kurtosis is usually compared to the kurtosis of a normal distribution, which is 3. Random variables with a kurtosis greater than 3 are often called heavy-tailed/fat-tailed, because the frequency of large deviations is higher than that of a random variable with a normal distribution.[8]

The estimators of the skewness and the kurtosis both use the principle of the sample analog to the expectation operator. The estimators for the skewness and kurtosis are

$$\frac{\hat{\mu}_3}{\hat{\sigma}^3} \text{ and } \frac{\hat{\mu}_4}{\hat{\sigma}^4} \quad (5.11)$$

[6] See Federal Reserve Bank of St. Louis. Federal Reserve Economic Data: FRED: St. Louis Fed. Retrieved from https://fred.stlouisfed.org/ These data are freely available.

[7] Thursday-to-Thursday prices are preferred when computing returns because public holidays are more likely to occur on a Friday.

[8] Many software packages report excess kurtosis, which is the kurtosis minus 3 (i.e., k – 3), and so care is needed when using reported kurtosis to assess tail heaviness.

Table 5.1 Sample Statistics of the Returns of Four Representative Financial Assets. The Returns Are Computed from Daily, Weekly (Using Thursday Prices), and End-of-Month Prices. Each Panel Reports the Mean, Standard Deviation, Skewness, and Kurtosis. The Annualized Mean and Standard Deviation Are Reported and Are Based on Scale Factors of 245, 52, and 12 for Daily, Weekly, and Monthly Data, Respectively. The Row Labeled *n* Reports the Number of Observations. The Final Row in the Top Panel Reports the Average Number of Observations per Year

	Daily Data			
	Stocks	**Bonds**	**Crude**	**Gold**
Mean	0.044%	0.027%	0.045%	0.021%
Std. Dev.	1.297%	0.301%	2.497%	1.002%
Skewness	−0.202	−0.321	0.092	−0.095
Kurtosis	6.39	2.62	12.78	6.56
Ann. Mean	10.7%	6.7%	10.9%	5.2%
Ann. Std. Dev.	20.3%	4.7%	39.1%	15.7%
n	7326	7326	7326	7326
n/Year	245	245	245	245
	Weely Data			
	Stocks	**Bonds**	**Crude**	**Gold**
Mean	0.205%	0.128%	0.191%	0.096%
Std. Dev.	2.747%	0.652%	5.119%	2.157%
Skewness	−0.676	−0.396	−0.074	0.211
Kurtosis	8.66	2.58	2.81	4.83
Ann. Mean	10.6%	6.7%	9.9%	5.0%
Ann. Std. Dev.	19.8%	4.7%	36.9%	15.6%
n	1561	1561	1561	1561
	Monthly Data			
	Stocks	**Bonds**	**Crude**	**Gold**
Mean	0.866%	0.559%	0.709%	0.427%
Std. Dev.	5.287%	1.462%	9.375%	4.460%
Skewness	−0.445	−0.748	0.277	0.147
Kurtosis	0.92	4.73	2.24	1.35
Ann. Mean	10.4%	6.7%	8.5%	5.1%
Ann. Std. Dev.	18.3%	5.1%	32.5%	15.4%
n	359	359	359	359

The third and fourth central moments are estimated by:

$$\hat{\mu}_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^3 \quad (5.12)$$

$$\hat{\mu}_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^4 \quad (5.13)$$

Table 5.1 also reports the estimated skewness and kurtosis for the four assets and three sampling horizons. These moments are scale-free, and so there is no need to annualize them.

Skewness, while being unit-free, does not follow a simple scaling law across sampling frequencies. Financial data are heteroskedastic

HISTOGRAMS AND KERNEL DENSITY PLOTS

Histograms are simple graphical tools used to represent the frequency distribution of a data series. Histograms divide the range of the data into *m* bins and then tabulate the number of values that fall within the bounds of each bin. Kernel density plots are smoothed versions of histogram plots.

A density plot differs from a histogram in two ways. First, rather than using discrete bins with well-defined edges, a density plot computes the number of observations that are close to any point on the x-axis. In effect, a density plot uses as many bins as there are data points in the sample. Second, a density plot uses a weighted count where the weight depends on the distance between the point on the x-axis and the observed value. Most common kernel density estimators use a weight function that declines as the distance between two points increases, thus producing a smooth curve.

Figure 5.2 shows two examples of histograms and kernel densities. The left panel shows the daily returns of the Russell 2000 Index, while the right panel shows its monthly returns. Because there are over 7,000 daily returns, the density plot is only slightly smoother than the histogram. In the monthly example, there are only 383 observations and so the difference between the two is more pronounced. Kernel density plots are generally preferred over histograms when visualizing the distribution of a dataset.

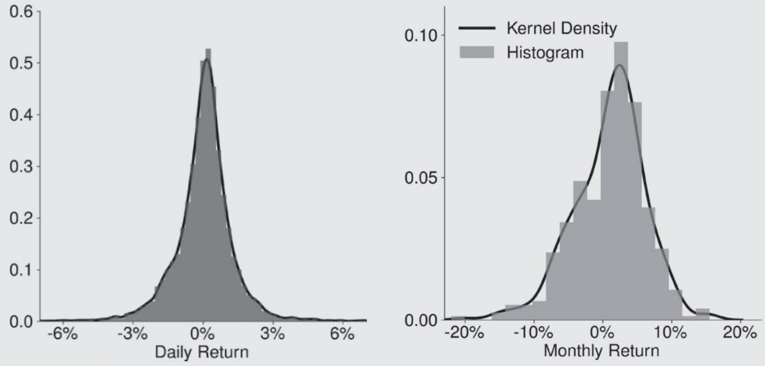


Figure 5.2 Histograms and densities for daily (left panel) and monthly (right panel) Russell 2000 Index returns.

(i.e., the volatility of return changes over time) and the differences in the volatility dynamics across assets produce a different scaling of skewness.[9] Gold has a positive skewness, especially at longer horizons, indicating that positive surprises are more likely than negative surprises. The skewness of crude oil returns is negative when returns are sampled at the daily horizon but becomes positive when sampling over longer horizons. Bond and stock market returns are usually negatively skewed at all horizons since large negative returns occur more frequently than comparably large positive returns.

All four asset return series have excess kurtosis, although bonds have fewer extreme observations than the other three asset classes. Kurtosis declines at longer horizons, and it is a stylized fact that returns sampled at low frequencies (i.e., monthly, or even quarterly) have a distribution that is closer to that of a normal than returns sampled at higher frequencies. This is because large short-term changes are diluted over longer horizons by periods of relative calmness.

Figure 5.3 contains density plots of the four financial assets in the previous example. Each plot compares the density of the assets' returns to that of a normal distribution with the same mean and variance. These four plots highlight some common features that apply across asset classes. The most striking feature of all four plots is the heavy tails. Note that the extreme tails (i.e., the values outside $\pm 2\sigma$) of the empirical densities are above those of the normal. The more obvious hint that these distributions are heavy-tailed is the sharp peak in the center of each density. Because both the empirical density and the normal density match the sample variance, the contribution of the heavy tails to the variance must be offset by an increase in mass

[9] Time-varying volatility is a pervasive property of financial asset returns. This topic is covered in Chapter 13.

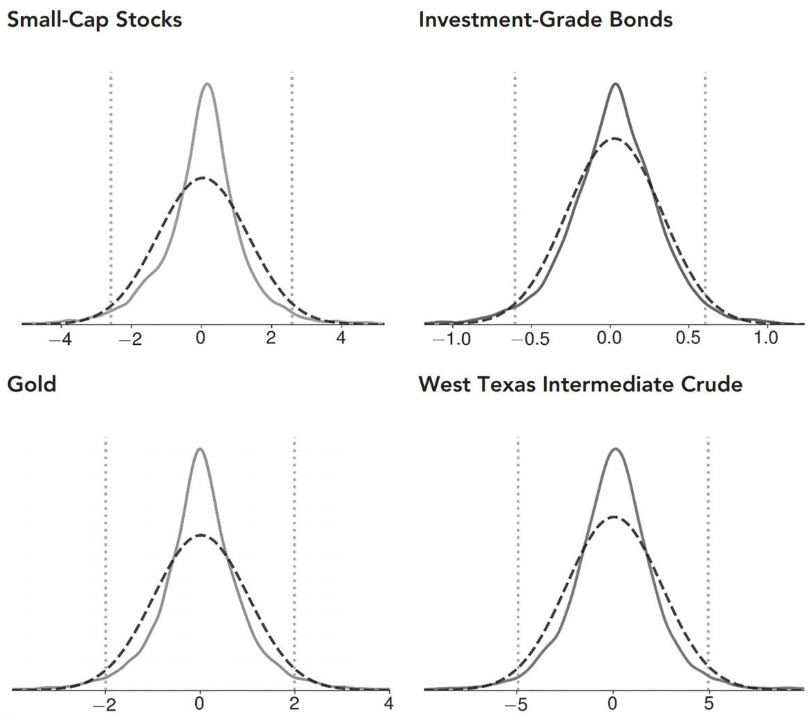


Figure 5.3 Density plots of daily returns data from small-cap stocks, corporate bonds, gold, and crude oil. The solid line in each plot is a data-based density estimate of each asset's returns. The dashed line in each panel is the density of a normal random variable with the same mean and variance as the corresponding asset returns. The vertical lines indicate $\pm 2\sigma$.

near the center. At the same time, regions near $\pm 1\sigma$ lose mass to the center and the tail to preserve the variance. The negative skewness is also evident in stock, bond, and crude oil returns. All three densities appear to be leaning to the right, which is evidence of a long left-tail (although the fat tails themselves are hard to see with figures of this size).

5.3 THE BLUE MEAN ESTIMATOR

The mean estimator is the Best Linear Unbiased Estimator (BLUE) of the population mean when the data are iid.[10] In this context, best indicates that the mean estimator has the lowest variance of any linear unbiased estimator (LUE).

[10] BLUE only requires that the mean and variance are the same for all observations. These conditions are satisfied when random variables are iid.

Linear estimators of the mean can be expressed as:

$$\hat{\mu} = \sum_{i=1}^{n} w_i X_i,$$

where w_i are weights that do not depend on X_i. In the sample mean estimator, for example, $w_i = 1/n$.

The unbiasedness of the mean estimator was shown earlier in this chapter. Showing that the mean is BLUE is also straightforward, but it is quite tedious and so is left as an exercise.[11]

BLUE is a desirable property for an estimator, because it establishes that the estimator is the best estimator (in the sense of

[11] To show that this claim is true, consider another linear estimator that uses a set of weights of the form $\tilde{w}_i = w_i + d_i$. Unbiasedness requires that $\sum \tilde{w}_i = 1$, and so $\sum d_i = 0$. The remaining steps only require computing the variance, which is equal to the variance of the sample mean estimator plus a term that is always positive.

having the smallest variance) among all linear and unbiased estimators. It does not, however, imply that there are no superior estimators to the sample mean. It only implies that these estimators must either be biased or nonlinear. Maximum likelihood estimators of the mean are generally more accurate than the sample mean, although they are usually nonlinear and often biased in finite samples. Maximum likelihood estimation will be discussed in detail in Chapter 6.

5.4 LARGE SAMPLE BEHAVIOR OF THE MEAN

The mean estimator is always unbiased, and the variance of the mean estimator takes a simple form when the data are iid. Two moments, however, are usually not enough to completely describe the distribution of the mean estimator. If data are iid normally distributed, then the mean estimator is also normally distributed.[12] However, it is generally not possible to establish the exact distribution of the mean based on a finite sample of n observations. Instead, modern econometrics exploits the behavior of the mean in large samples (formally, when $n \to \infty$) to approximate the distribution of the mean in finite samples.

The Law of Large Numbers (LLN) establishes the large sample behavior of the mean estimator and provides conditions where an average converges to its expectation. There are many LLNs, but the simplest for iid random variables is the Kolmogorov Strong Law of Large Numbers. This LLN states that if $\{X_i\}$ is a sequence of iid random variables with $\mu \equiv E[X_i]$, then:

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{a.s.} \mu$$

The symbol $\xrightarrow{a.s.}$ means *converges almost surely.*[13] This property ensures that the probability of $\hat{\mu}_n$ being far from μ converges to 0 as $n \to \infty$.

When an LLN applies to an estimator, the estimator is said to be *consistent*. Consistency requires that an estimator is asymptotically unbiased, and so any finite sample bias must diminish as n increases (e.g., $\hat{\sigma}^2$ has this property). A consistent estimator, however, has another key feature: as the sample size n grows larger, the variance of the estimator converges to zero (i.e., $V[\hat{\mu}_n] \to 0$). This ensures that the chance of observing a large deviation of a sample estimate from the population value is negligible in large samples.

Figure 5.4 illustrates the LLN using simulated data. The figures on the left all use log-normal data with parameters $\mu = 0$ and $\sigma^2 = 1$. The figures on the right use data generated by a Poisson with shape parameter equal to 3. Both distributions are right-skewed, and the Poisson is a discrete distribution.

The top panels show the PDF and the PMF of the simulated data used in the illustration. The middle panel shows the distribution of the sample mean using simulated data. The number of simulated data points varies from 10 to 640 in multiples of 4. This ratio between the sample sizes ensures that the standard deviation halves each time the sample size increases. The dashed line in the center of the plots shows the population mean. These finite-sample distributions of the mean estimators are calculated using data simulated from the assumed distribution. The sample mean is computed from each simulated sample, and 10,000 independent samples are constructed. The plots in the center row show the empirical density of the estimated sample means for each sample size.

Two features are evident from the LLN plots. First, the density of the sample mean is not a normal. In sample means computed from the simulated log-normal data, the distribution of the sample mean is right-skewed, especially when n is small. Second, the distribution becomes narrower and more concentrated around the population mean as the sample size increases. The collapse of the distribution around the population mean is evidence that the LLN applies to these estimators.

The LLN also applies to other sample moments. For example, when the data are iid and $E[X_i^2]$ is finite, then the LLN ensures that $\hat{\sigma}^2 \xrightarrow{a.s.} \sigma^2$, and so the sample variance estimator is also consistent.[14]

Consistency is an important property of an estimator, although it is not enough to understand the estimator's distribution. For example, the distribution of $\hat{\mu} - \mu$ is not easy to study because it collapses to 0 as $n \to \infty$.

The solution is to rescale the difference between the estimate and population value by $\sqrt{n}$. This is the required value to stabilize the distribution because

$$V[\hat{\mu}] = \frac{\sigma^2}{n}$$

so that

$$V[\sqrt{n}\hat{\mu}] = nV[\hat{\mu}] = \sigma$$

The Centrl Limit Theorem (CLT) states that the sampling distribution of the mean for any random sample of observations will:

- tend towards the normal distribution, and
- have a mean equal to the population mean as the sample size tends to infinity.

[12] This result only depends on the property that the sum of normal random variables is also normally distributed.

[13] Almost sure convergence is a technical concept from probability theory. An estimator that converges almost surely must converge for any sequence of values that can be produced by the random variables X_i, $i = 1, 2, \ldots$.

[14] Formally the LLN can be applied to two averages, $n^{-1}\sum x_i^2 \xrightarrow{a.s.} E[X^2]$ and $n^{-1}\sum x_i \xrightarrow{a.s.} E[X]$. The sample variance estimator can be equivalently expressed using these two averages.

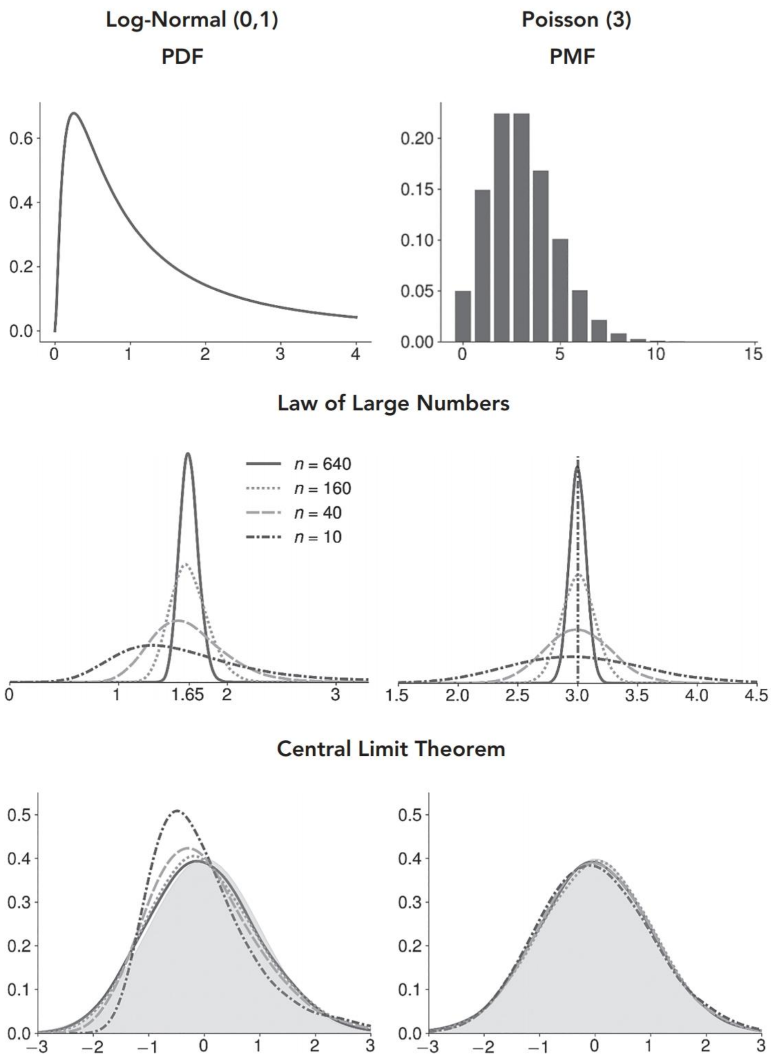


Figure 5.4 These figures demonstrate the consistency of the sample mean estimator using the Law of Large Numbers (middle panels) and with respect to the normal distribution when scaled (bottom panels). The figures are generated using simulated data. The left panels use simulated data from a log-normal distribution with parameters $\mu = 0, \sigma^2 = 1$. The right panels use simulated data from a discrete distribution, the Poisson with shape parameter equal to 3. The top panels show the PDF of the log-normal and the PMF of the Poisson.

Note that the CLT can also be applied to the rescaled difference under some additional assumptions.

The simplest CLT is known as the Lindberg-Lévy CLT. This CLT states that if $\{X_i\}$ are iid, then

$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} N(0, \sigma^2),$$

where $\mu = E[X_i]$ and $\sigma^2 = V[X_i]$. The symbol $\xrightarrow{d}$ denotes convergence in distribution.

This CLT requires a further assumption in addition to those required for the LLN: that the variance σ^2 is finite (as the LLN only requires that the mean is finite). This CLT can be alternatively expressed as:

$$\sqrt{n}\left(\frac{\hat{\mu} - \mu}{\sigma}\right) = \left(\frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}}\right) \xrightarrow{d} N(0,1)$$

This version makes it clear that the mean, when scaled by its standard error (i.e., $\sigma/\sqrt{n}$), is asymptotically standard normally distributed.

CLTs extend LLNs and provide an approximation to the distribution of the sample mean estimator. Furthermore, they do not require knowledge of the distribution of the random variables generating the data. In fact, only independence and some moment conditions are required for the CLT to apply to a sample mean estimator.

The CLT is an asymptotic result and so technically only holds asymptotically (i.e., in the limit). In practice, the CLT is used as an approximation in finite samples so that the distribution of the sample mean is approximated by:

$$\hat{\mu} \sim N\left(\mu, \sigma^2/n\right)$$

This expression for the mean shows that (in large samples) the distribution of the sample mean estimator ($\hat{\mu}$) is centered on the population mean (μ), and the variance of the sample average declines as n grows.

The bottom panels of Figure 5.4 show the empirical distribution for 10,000 simulated values of the sample mean. The CLT says that the sample mean is normally distributed in large samples, so that $\hat{\mu} \sim N(\mu, \sigma^2/n)$. This value is standardized by subtracting the mean and dividing by the standard errors of the mean so that:

$$Z = \frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}}$$

is a standard normal random variable. The simulated values show whether the CLT provides an accurate approximation to the (unknown) true distribution of the sample mean for different sample sizes n.

Determining when n is large enough is *the* fundamental question when applying a CLT. Returning to Figure 5.4, note that the shaded areas in the bottom panels are the PDF of a standard normal. For data generated from the log-normal (i.e., the left panel), the CLT is *not* an accurate approximation when $n = 10$ because of the evident right-skewness (due to the skewness in the random data sampled from the log-normal). When $n = 40$, however, the skew has substantially disappeared and the approximation by a standard normal distribution appears to be accurate. Note that skewness in the distribution of the sample mean is a finite-sample property, and when $n = 160$, the distribution is extremely close to the standard normal. This shows that the sample mean estimator is asymptotically normally distributed even if the underlying data follow some other (non-normal) distribution.

In the application of the CLT to the mean of data generated from Poisson random variables (i.e., the right panel), the CLT appears to be accurate for all sample sizes.

5.5 THE MEDIAN AND OTHER QUANTILES

The median, like the mean, measures the central tendency of a distribution. Specifically, it is the 50% quantile of the distribution and the point where the probabilities of observing a value above or below it are equal (i.e., 50%). When a distribution is symmetric, the median is in the center of the distribution and is the same as the mean. When distributions are asymmetric, the median is larger (smaller) than the mean if the distribution is left-skewed (right-skewed).

Estimation of the median from a data sample is simple. First, the data are sorted from smallest to largest. When the sample size is odd, the value in position $(n + 1)/2$ of the sorted list is used to estimate the median:

$$\text{median}(x) = x_{(n+1)/2} \tag{5.14}$$

When the sample size is even, the median is estimated using the average of the two central points of the sorted list:

$$\text{median}(x) = (1/2)(x_{n/2} + x_{n/2+1}). \tag{5.15}$$

Two other commonly reported quantiles are the 25% and 75% quantiles. These are estimated using the same method as the median. More generally, the α-quantile is estimated from the sorted data using the value in location αn. When αn is not an integer value, then the usual practice is to take the average of the points immediately above and below αn.[15]

These two quantile estimates ($\hat{q}_{25}$ and $\hat{q}_{75}$) are frequently used together to estimate the inter-quartile range:

$$\widehat{IQR} = \hat{q}_{75} - \hat{q}_{25}$$

The *IQR* is a measure of dispersion and so is an alternative to the standard deviation. Other quantile measures can be constructed to measure the extent to which the series is asymmetric or to estimate the heaviness of the tails.

[15] There is a wide range of methods to interpolate quantiles when αn is not an integer, and different methods are used across the range of common statistical software.

Two features make quantiles attractive. First, quantiles have the same units as the underlying data, and so they are easy to interpret. For example, the 25% quantile for a sample of asset returns estimates the point where there is 25% probability of observing a smaller return (and a 75% probability of observing a larger return). Meanwhile, the interquartile range estimates a central interval where there is 50% probability of observing a return.

The second attractive feature is robustness to outliers (i.e., values very far from the mean). For example, suppose that some observed data $\{x_1, \ldots, x_n\}$ are contaminated with an outlier. Both the median and *IQR* are unaffected by the presence of an outlier. However, this is not true of the mean estimator, which gives weight $1/n$ to the outlier. If the outlier is far from the other (valid) observations, this distortion can be large. The variance, because it squares the difference between the outlier and the mean, is even more sensitive to outliers.

Table 5.2 contains the median, quantile, and IQR estimates for the four-asset example discussed in this chapter. The median return is above the mean return for the negatively skewed assets (i.e., bonds, stocks, and crude oil). Note that the median scales similarly to the mean, and the weekly median is about five times larger than the daily value.

Table 5.2 Quantile Estimates for the Four Asset Classes Using Data Sampled Daily, Weekly, and Monthly. The Median Is the 50% Quantile. The Inter-Quartile Range (IQR) Is the Difference between the 75% Quantile q(.75) and the 25% Quantile q(.25)

	Daily Data			
	Bonds	**Stocks**	**Gold**	**Crude**
Median	0.034%	0.103%	0.014%	0.058%
q(.25)	−0.140%	−0.534%	−0.445%	−1.21%
q(.75)	0.207%	0.660%	0.496%	1.30%
IQR	0.347%	1.19%	0.941%	2.51%
	Weekly Data			
	Bonds	**Stocks**	**Gold**	**Crude**
Median	0.159%	0.459%	0.044%	0.224%
q(.25)	−0.244%	−1.12%	−1.01%	−2.64%
q(.75)	0.542%	1.69%	1.24%	3.03%
IQR	0.786%	2.81%	2.24%	5.67%
	Monthly Data			
	Bonds	**Stocks**	**Gold**	**Crude**
Median	0.640%	1.56%	0.265%	0.839%
q(.25)	−0.220%	−2.47%	−2.29%	−5.14%
q(.75)	1.42%	4.12%	2.84%	6.72%
IQR	1.64%	6.59%	5.14%	11.9%

The interquartile range also agrees with the standard deviation estimates; crude is the riskiest of the assets, followed by stocks and gold, while bonds are the safest (assuming that the risk of each asset can be measured by the dispersion of its returns).

5.6 MULTIVARIATE MOMENTS

The sample analog can be extended from univariate statistics (i.e., involving a single random variable) to multivariate statistics (i.e., involving two or more random variables).

The extension of the mean is straightforward. The sample mean of two series is just the collection of the two univariate sample means. However, extending the sample variance estimator to multivariate datasets requires estimating the variance of each series and the covariance between each pair. The higher-order moments, such as skewness and kurtosis, can be similarly defined in a multivariate setting.

Covariance and Correlation

Recall that covariance measures the linear dependence between two random variables and is defined as:

$$\sigma_{XY} \equiv \text{Cov}[X,Y] = \text{E}[(X - \text{E}[X])(Y - \text{E}[Y])]$$

The sample covariance estimator uses the sample analog to the expectation operator

$$\hat{\sigma}_{XY} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_X)(Y_i - \hat{\mu}_Y), \tag{5.16}$$

where $\hat{\mu}_X$ is the sample mean of X and $\hat{\mu}_Y$ is the sample mean of Y. Like the sample variance estimator, the sample covariance estimator is biased toward zero. Dividing by $n - 1$, rather than n, produces an unbiased estimate of the covariance.

Correlation is the standardized version of the covariance and is usually preferred because it does not depend on the scale of X or Y. The correlation is estimated by dividing the sample covariance by the product of the sample standard deviations of each variable:

$$\hat{\rho}_{XY} = \frac{\hat{\sigma}_{XY}}{\sqrt{\hat{\sigma}_X^2}\sqrt{\hat{\sigma}_Y^2}} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X\hat{\sigma}_Y} \tag{5.17}$$

The sample correlation estimator is biased in finite samples, even if unbiased estimators are used to estimate the variances and covariance. This occurs for the same reason that the sample estimator of the standard deviation is always biased: Unbiasedness is not preserved through nonlinear transformations (because $\text{E}[g(Z)] \neq g(\text{E}[Z])$ for nonlinear functions g).

Table 5.3 contains the estimated correlations between the returns in the four-asset example. These correlations are estimated using returns computed from daily, weekly, and monthly data. Note that correlations are unit free by construction.

The highest correlation (between gold and crude oil) ranges between 13% and 19% across the different sampling frequencies. Stocks and bonds have a negative correlation when measured using daily data and a positive correlation when measured using monthly data. Correlations measured at different frequencies do not have to be the same, and here the monthly correlations are generally larger than the daily correlations. This is because short-term returns are often driven by short-term issues that are specific to an individual asset (e.g., liquidity), whereas longer-term returns are more sensitive to macroeconomic changes.

These are all relatively small correlations, and cross-asset class correlations are generally smaller than correlations within an asset class. For example, the estimate of the correlation between gold and stocks is small (i.e., between −2 and 4%). In contrast, the sample correlation between the Russell 1000 (a large-cap stock index) and the Russell 2000 (a small-cap stock index) is 87%, 88%, and 85% when measured using daily, weekly, or monthly data, respectively. This feature is useful when constructing diversified portfolios, because the correlation between the assets held appears in the variance of the portfolio.

Table 5.3 **Sample Correlations between Bond, Stock, Gold, and Crude Oil Returns. The Correlations Are Measured Using Data Sampled Daily, Weekly, and Monthly**

	Daily Data			
	Bonds	**Stocks**	**Gold**	**Crude**
Bonds	–	−16.4%	3.0%	−9.9%
Stocks	−16.4%	–	−2.1%	12.5%
Gold	3.0%	−2.1%	–	13.0%
Crude	−9.9%	12.5%	13.0%	–
	Weekly Data			
	Bonds	**Stocks**	**Gold**	**Crude**
Bonds	–	−2.4%	2.3%	−7.7%
Stocks	−2.4%	–	4.0%	13.2%
Gold	2.3%	4.0%	–	15.2%
Crude	−7.7%	13.2%	15.2%	–
	Monthly Data			
	Bonds	**Stocks**	**Gold**	**Crude**
Bonds	–	15.1%	19.6%	−0.2%
Stocks	15.1%	–	−0.2%	11.6%
Gold	19.6%	−0.2%	–	18.9%
Crude	−0.2%	11.6%	18.9%	–

Sample Mean of Two Variables

Estimating the means of two random variables is no different from estimating the mean of each separately, so that:

$$\hat{\mu}_X = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ and } \hat{\mu}_y = \frac{1}{n}\sum_{i=1}^{n} y_i$$

When data are iid, the CLT applies to each estimator. However, it is more useful to consider the joint behavior of the two mean estimators by treating them as a bivariate statistic. The CLT can be applied by stacking the two mean estimators into a vector:

$$\hat{\mu} = \begin{bmatrix} \hat{\mu}_X \\ \hat{\mu}_Y \end{bmatrix} \quad (5.18)$$

This 2-by-1 vector is asymptotically normally distributed if the multivariate random variable $Z = [X, Y]$ is iid.[16]

The CLT for the vector depends on the 2-by-2 covariance matrix for the data. A covariance matrix collects the two variances (one for X and one for Y) and the covariance between X and Y into a matrix:

$$\begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix} \quad (5.19)$$

The elements along the leading diagonal of the covariance matrix are the variances of each series, and those on the off-diagonal are the covariance between the pair.

The CLT for a pair of mean estimators is virtually identical to that of a single mean estimator, where the scalar variance is replaced by the covariance matrix. The CLT for bivariate iid data series states that:

$$\sqrt{n}\begin{bmatrix} \hat{\mu}_X - \mu_X \\ \hat{\mu}_Y - \mu_Y \end{bmatrix} \rightarrow N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix}\right) \quad (5.20)$$

so that the scaled difference between the vector of means is asymptotically *multivariate* normally distributed. In practice, this CLT is applied by treating the mean estimators as a bivariate normal random variable:

$$\begin{bmatrix} \hat{\mu}_X \\ \hat{\mu}_Y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \frac{\sigma_X^2}{n} & \frac{\sigma_{XY}}{n} \\ \frac{\sigma_{XY}}{n} & \frac{\sigma_Y^2}{n} \end{bmatrix}\right)$$

[16] This assumes that each component has a finite variance.

Table 5.4 Estimated Moments of the Monthly Returns on the Russell 2000 and the BoAML Total Return Index During the Period between 1987 and 2018. The Means, Variance, and Covariance Are All Annualized. The Correlation Is Scale-Free

Moment	$\hat{\mu}_S$	$\hat{\sigma}_S^2$	$\hat{\mu}_B$	$\hat{\sigma}_B^2$	σ_{SB}	ρ
	10.4	335.4	6.71	25.6	14.0	0.151

Table 5.4 presents the annualized estimates of the means, variances, covariance, and correlation for the monthly returns on the Russell 2000 and the BoAML Total Return Index (respectively symbolized by *S* and *B*). Note that the returns on the Russell 2000 are much more volatile than the returns on the corporate bond index. The correlation between the returns on these two indices (i.e., 0.151) is positive but relatively small.

The CLT depends on the population values for the two variances (σ_S^2 and σ_B^2) and the covariance between the two ($\sigma_{SB} = \rho\sigma_S\sigma_B$). To operationalize the CLT, the population values are replaced with estimates computed using the sample variance of each series and the sample covariance between the two series. Applying the bivariate CLT using the numbers in Table 5.4:

$$\sqrt{n}\begin{bmatrix}\hat{\mu}_S - \mu_S \\ \hat{\mu}_B - \mu_B\end{bmatrix} \rightarrow N\left(\begin{bmatrix}0 \\ 0\end{bmatrix}, \begin{bmatrix}335.4 & 14.0 \\ 14.0 & 25.6\end{bmatrix}\right)$$

Recall that in practice, the definition is applied as if the mean estimators follow the limiting distribution, and so the estimates are treated as if they are normally distributed:

$$\begin{bmatrix}\hat{\mu}_S \\ \hat{\mu}_B\end{bmatrix} \sim N\left(\begin{bmatrix}\mu_S \\ \mu_B\end{bmatrix}, \begin{bmatrix}0.934 & 0.039 \\ 0.039 & 0.071\end{bmatrix}\right),$$

where the covariance matrix is equal to the previous covariance matrix divided by *n* (which is 359 in this example).

The most important insight from the bivariate CLT is that correlation in the data produces a correlation between the sample means. Moreover, the correlation between the means is identical to the correlation between the data series.

Coskewness and Cokurtosis

Like variance, skewness and kurtosis can be extended to pairs of random variables. When computing cross p^{th} moments, there are $p - 1$ different measures. Applying the principle to the first four moments there are

- No cross means,
- One cross variance (covariance),
- Two measures of cross-skewness (coskewness), and
- Three cross-kurtoses (cokurtosis).

The two coskewness measures are

$$s(X,X,Y) = \frac{E[(X - E[X])^2(Y - E[Y])]}{\sigma_X^2\sigma_Y}$$

and

$$s(X,Y,Y) = \frac{E[(X - E[X])(Y - E[Y])^2]}{\sigma_X\sigma_Y^2} \qquad (5.21)$$

The coskewness standardizes the cross-third moments by the variance of one of the variables and the standard deviation of the other, and so is scale- and unit-free.

These measures both capture the likelihood of the data taking a large directional value whenever the other variable is large in magnitude. When there is no sensitivity to the direction of one variable to the magnitude of the other, the two coskewnesses are 0. For example, the coskewness in a bivariate normal is always 0, even when the correlation is different from 0. Written using the same notation as equation (5.21), the univariate skewness estimators are $s(X,X,X)$ and $s(Y,Y,Y)$.

Coskewnesses are estimated by applying the sample analog to the expectation operator to the definition of coskewness. For example:

$$\hat{s}(X,X,Y) = \frac{1/n\sum_{i=1}^{n}(x_i - \hat{\mu}_X)^2(y_i - \hat{\mu}_Y)}{\hat{\sigma}_X^2\hat{\sigma}_Y}$$

Table 5.5 contains the skewness and coskewness for the four assets (i.e., bonds, stocks, crude oil, and gold) using data sampled monthly. The estimates of the coskewness are mostly negative. The bond-stock and crude oil-stock pairs have the largest coskewness, although neither relationship is symmetric. For example, $s(X,X,Y)$ in the bond-stock pair indicates that stock returns tend to be negative when bond volatility is high. However, $s(X,Y,Y)$ is close to 0 so that the sign of bond returns does not appear to be strongly linked to the volatility of stock returns.

Table 5.5 Skewness and Coskewness in the Group of Four Asset Classes Estimated Using Monthly Data. The Far Left and Far Right Columns Contain Skewness Measures of the Variables Labeled *X* and *Y*, Respectively. The Middle Two Columns Report the Estimated Coskewness

X	Y	s(X,X,X)	s(X,X,Y)	s(X,Y,Y)	s(Y,Y,Y)
Bonds	Crude	−0.179	−0.008	0.040	0.104
Bonds	Gold	−0.179	−0.018	−0.032	−0.101
Bonds	Stocks	−0.179	−0.082	0.012	−0.365
Crude	Gold	0.104	−0.010	−0.098	−0.101
Crude	Stocks	0.104	−0.064	−0.127	−0.365
Gold	Stocks	−0.101	−0.005	0.014	−0.365

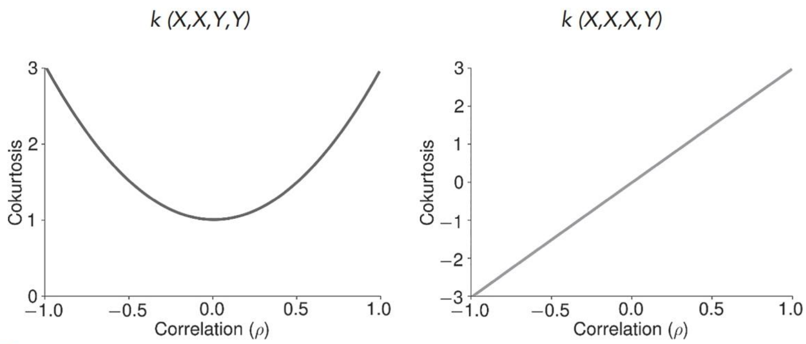


Figure 5.5 Plots of the cokurtoses for the symmetric case ($\kappa(X,X,Y,Y)$, left) and the asymmetric case ($\kappa(X,Y,Y,Y)$, right). Note that $\kappa(X,Y,Y,Y)$ has the same shape as $\kappa(X,X,X,Y)$ and so is omitted.

The cokurtosis uses combinations of powers that add to 4, and so there are three configurations: (1,3), (2,2), and (3,1). The (2,2) measure is the easiest to understand, and captures the sensitivity of the magnitude of one series to the magnitude of the other series (i.e., the strength of the relationship between the volatility of X and the volatility of Y). If both series tend to be large in magnitude at the same time, then the (2,2) cokurtosis is large. The other two kurtosis measures, (3,1) and (1,3), capture the agreement of the return signs when the power 3 return is large.

The three measures of cokurtosis are defined as:

$$\kappa(X,X,Y,Y) = \frac{E[(X - E[X])^2(Y - E[Y])^2]}{\sigma_X^2\sigma_Y^2} \quad (5.22)$$

$$\kappa(X,X,X,Y) = \frac{E[(X - E[X])^3(Y - E[Y])]}{\sigma_X^3\sigma_Y} \quad (5.23)$$

$$\kappa(X,Y,Y,Y) = \frac{E[(X - E[X])(Y - E[Y])^3]}{\sigma_X\sigma_Y^3} \quad (5.24)$$

When examining kurtosis, the value is usually compared to the kurtosis of a normal distribution (which is 3). Comparing a cokurtosis to that of a normal is more difficult, because the cokurtosis of a bivariate normal depends on the correlation.

Figure 5.5 contains plots of the cokurtoses for normal data as a function of the correlation between the variables (i.e., ρ), which is always between −1 and 1. The symmetric cokurtosis $\kappa(X,X,Y,Y)$ ranges between 1 and 3: it is 1 when returns are uncorrelated (and therefore independent, because the two random variables are normal) and rises symmetrically as the correlation moves away from 0. The asymmetric cokurtosis ranges from −3 to 3 and is linear in ρ.

Table 5.6 contains estimates of the kurtosis and cokurtosis for the four assets previously described. The sample kurtosis is computed using the sample analog to the expectation operator. The first and last columns contain the kurtosis of the two assets

Table 5.6 Kurtosis and Cokurtosis in the Group of Four Asset Classes Estimated Using Monthly Data. The Far Left and Far Right Columns Contain Kurtosis Measures of the Variables Labeled *X* and *Y*, Respectively. The Middle Three Columns Report the Three Cokurtosis Estimates

X	Y	κ(X,X,X,X)	κ(X,X,X,Y)	κ(X,X,Y,Y)	κ(X,Y,Y,Y)	κ(Y,Y,Y,Y)
Bonds	Crude	6.662	−0.248	1.996	0.244	15.656
Bonds	Gold	6.662	−0.537	1.808	0.547	9.390
Bonds	Stocks	6.662	−1.522	2.921	0.682	11.020
Crude	Gold	15.656	4.010	2.833	7.011	9.390
Crude	Stocks	15.656	−0.443	2.490	4.750	11.020
Gold	Stocks	9.390	−0.081	3.230	2.491	11.020

reported in the row. The center column reports the symmetric cokurtosis $\kappa(X,X,Y,Y)$, which measures the strength of the link between the variances of the two series. Note that gold and stocks appear to have the strongest link.

The two columns $\kappa(X,X,X,Y)$ and $\kappa(X,Y,Y,Y)$ measure the strength of the dependence in the signed extremes. In each column, the extreme values of the variable that appears three times have more influence on the cokurtosis. For example, the bond-stock pair $\kappa(X,X,X,Y)$ is -1.5, which indicates that the return on stocks tends to have the opposite sign of the return on bonds when bonds are in their tail. Meanwhile, $\kappa(X,Y,Y,Y)$ is positive, which indicates that the bond returns tend to have the same sign as the stock returns when stocks have an extreme observation. Additionally, all of the $\kappa(X,X,Y,Y)$ are positive as well, indicating that (on average) where volatility is high in one market, it is also high in the other.

5.7 SUMMARY

This chapter describes the steps required to estimate four important moments: the mean, the variance, the skewness, and the kurtosis. These statistics are widely reported to describe observed data, and each measures an important feature in the distribution of a random variable. The key properties of the mean estimator include the following:

- The mean is unbiased.
- When the observed data are iid, the mean has a simple expression for its standard error.
- The mean estimator is BLUE.
- The mean estimator is consistent, and in large samples the estimated mean is close to the population mean.
- When the observed data are iid and the variance is finite, the distribution of the mean estimator can be approximated using the CLT.

All moment estimators use the principle of the sample analog, where the expectation operators in the population moment are replaced by the sample average. The skewness describes the symmetry of the data and the kurtosis describes the heaviness of the tails. Data series representing four major asset classes—bonds, equities, gold, and energy—all confirm two stylized facts of financial return data: The distributions of financial asset returns are not symmetric (i.e., they are skewed) and heavy-tailed (i.e., have a kurtosis larger than that of a normal random variable).

Estimators based on quantiles (i.e., the median and the quartiles) complement sample moments. These values are simple to estimate and can be used to construct alternative measures to describe a data set. Quantile-based statistics are naturally robust to outliers and so are particularly attractive when using financial data.

The chapter concludes by introducing multivariate moment estimators. The sample covariance and correlation measure linear dependence between two data series. The CLT extends naturally to the mean of a bivariate random variable. The resulting bivariate CLT depends on the covariance matrix, which summarizes both the variances of two random variables and the covariance between them. Finally, the bivariate extensions of skewness and kurtosis—coskewness and cokurtosis—provide additional tools to describe the joint distribution of two random variables.

The following questions are intended to help candidates understand the material. They are *not* actual FRM exam questions.

QUESTIONS

Short Concept Questions

5.1 What is the difference between a sample mean and the population mean?

5.2 What is the difference between the standard deviation of the sample data and the standard error of the sample mean?

5.3 Is an unbiased estimator always better than a biased estimator?

5.4 What is the sample analog to the expectation operator, and how is it used?

5.5 Why is annualization useful, and how are means and standard deviations annualized?

5.6 What do skewness and kurtosis measure?

5.7 What is the skewness and kurtosis of a normal random variable?

5.8 How are quantiles estimated?

5.9 What advantages do quantiles have compared to moments?

5.10 When dealing with two random variables, how is coskewness interpreted?

Practice Questions

5.11 Suppose that four independent random variables X_1, X_2, X_3, and X_4 all have mean $\mu = 1$ and variances of $^1/_2$, $^1/_2$, 2, and 2, respectively. What is the expectation and variance of $\frac{1}{4}\sum X_i$?

5.12 Using the same information from question 5.11, compute the expectation and variance of $2X_1 + 2X_2 + \frac{1}{2}X_3 + \frac{1}{2}X_4$.

5.13 Using the same information from question 5.11, compute the expectation and variance of $\frac{2}{5}X_1 + \frac{2}{5}X_2 + \frac{1}{10}X_3 + \frac{1}{10}X_4$?

5.14 What is the effect on the sample covariance between X and Y if X is scaled by a constant a? What is the effect on the sample correlation?

5.15 An experiment yields the following data:

Trial Number	Value
1	0.00
2	0.07
3	0.13
4	0.13
5	0.20
6	0.23
7	0.25
8	0.27
9	0.34
10	0.41
11	0.60
12	0.66
13	0.76
14	0.77
15	0.96

It is hypothesized that the data come from a uniform distribution, $U(0, b)$.

a. Calculate the sample mean and variance.

b. What are the unbiased estimators of the mean and variance?

c. Calculate the b in $U(0, b)$ using the formula for the mean of a uniform distribution and the value of the unbiased sample mean found in part b.

d. Calculate the b in $U(0, b)$ using the formula for the variance of a uniform distribution and the value of the unbiased sample variance found in part b.

5.16 For the following data:

Observation	Value
1	0.38
2	0.28
3	0.27
4	0.99
5	0.26
6	0.43

a. What is the median?

b. What is the IQR?

The following questions are intended to help candidates understand the material. They are *not* actual FRM exam questions.

ANSWERS

Short Concept Questions

5.1 The sample mean is a numerical average of observed data. The population mean is the unknown (true) value that the sample mean is used to approximate.

5.2 When data are iid, the standard error of the mean is $\frac{\sigma}{\sqrt{n}}$, where σ is the standard deviation of the data. The standard error measures the estimation error in the sample mean. As the sample size grows the standard error converges to 0. The (population) standard deviation measures the dispersion in the data. It is a fixed value independent of the sample size.

5.3 Not necessarily. While an unbiased estimator is always right, on average, it may be very inaccurate, which would be reflected by a large variance. It may be better to use an estimator with a small bias but with lower variance. This tradeoff may be especially valuable if the bias depends on the sample size, decreasing as the sample size increases, and the sample size is large.

5.4 The sample analog replaces the expectation operator with the average so that $E[(g(X))]$ is approximated by $n^{-1}\sum g(x_i)$. The sample analog is a universally used approach to estimating moments of any order.

5.5 Annualization transforms moments measured using data sampled at any frequency—daily, weekly, monthly, quarterly—to a value that is equivalent under an assumption to a measure computed from annual data. The common scale factors are 252, 52, 12, and 4 to convert from daily, weekly, monthly, and quarterly. These scale factors are applied to the high-frequency mean or variance to produce annualized means and variances. The standard deviation is scaled by the square root of the scale.

5.6 Skewness measures the tendency to see larger values in one direction or the other relative to the mean. For example, a distribution with negative skew will tend to produce larger in magnitude values less than the mean. Kurtosis is a measurement of how frequently large observations of either sign are measured.

5.7 The skewness of a normal is 0, and the kurtosis is 3.

5.8 Quantiles are estimated by sorting the data from smallest to largest and then choosing the values with the sorted index $n\alpha$ to measure the quantile.

5.9 There are two potential advantages. First, quantiles are always well defined (exist) even when random variables are very heavy-tailed. Second, quantiles are less sensitive to outliers than moments.

5.10 Coskewness measures the likelihood of one variable taking a large directional value whenever the other variable is large in magnitude (i.e., it captures the likelihood that one variable is large when the other has a high variance). When there is no sensitivity of the direction of one variable to the magnitude of the other, the coskewness measures are 0.

Solved Problems

5.11 The expectation is $E\left[\frac{1}{4}(X_1 + X_2 + X_3 + X_4)\right] = \frac{1}{4}(E[X_1] + E[X_2] + E[X_3] + E[X_4]) = \frac{1}{4}(\mu + \mu + \mu + \mu) = \mu = 1.$

The variance is $Var\left[\frac{1}{4}(X_1 + X_2 + X_3 + X_4)\right] = \frac{1}{16}(Var[X_1] + Var[X_2] + Var[X_3] + Var[X_4]) = \frac{1}{16}\left(\frac{1}{2} + \frac{1}{2} + 2 + 2\right) = \frac{5}{16}.$

Note that the covariance terms are zero because the four variables are independent.

5.12 The expectation is $E\left[2X_1 + 2X_2 + \frac{1}{2}X_3 + \frac{1}{2}X_4\right] = 2\mu + 2\mu + \frac{1}{2}\mu + \frac{1}{2}\mu = 5\mu = 5.$ The variance is $Var\left[2X_1 + 2X_2 + \frac{1}{2}X_3 + \frac{1}{2}X_4\right] = 4Var[X_1] + 4Var[X_2] + \frac{1}{4}Var[X_3] + \frac{1}{4}Var[X_4] = \frac{4}{2} + \frac{4}{2} + \frac{2}{4} + \frac{2}{4} = 5.$

5.13 The expectation is computed using the same steps:

$$E\left[\frac{2}{5}X_1 + \frac{2}{5}X_2 + \frac{1}{10}X_3 + \frac{1}{10}X_4\right]$$
$$= \frac{2}{5}E[X_1] + \frac{2}{5}E[X_2] + \frac{1}{10}E[X_3] + \frac{1}{10}E[X_4]$$
$$= \frac{2}{5}\mu + \frac{2}{5}\mu + \frac{1}{10}\mu + \frac{1}{10}\mu = \mu = 1.$$

This estimator is unbiased.

The variance of the sum is the sum of the variances because the random variables are independent.

$$Var\left[\frac{2}{5}X_1 + \frac{2}{5}X_2 + \frac{1}{10}X_3 + \frac{1}{10}X_4\right]$$
$$= \frac{4}{25}Var[X_1] + \frac{4}{25}Var[X_2] + \frac{1}{100}Var[X_3] + \frac{1}{100}Var[X_4]$$
$$= \frac{4}{25} \times \frac{1}{2} + \frac{4}{25} \times \frac{1}{2} + \frac{1}{100} \times 2 + \frac{1}{100} \times 2 = \frac{5}{25} = \frac{1}{5}.$$

5.14 The sample covariance is defined $\frac{1}{n}\sum_{i=1}(X_i - \bar{X})(Y_i - \bar{Y})$. Multiplying each X_i by a constant a will rescale the sample mean by a so that $\frac{1}{n}\sum_{i=1}(aX_i - a\bar{X})(Y_i - \bar{Y}) = \frac{1}{n}\sum_{i=1}a(X_i - \bar{X})(Y_i - \bar{Y}) = a\frac{1}{n}\sum_{i=1}(X_i - \bar{X})(Y_i - \bar{Y})$. The final covariance is a times the original covariance. The correlation is then $\hat{\rho} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X\hat{\sigma}_Y}$ so that scaling X_i by a will produce $\hat{\rho} = \frac{a\hat{\sigma}_{XY}}{a\hat{\sigma}_X\hat{\sigma}_Y} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X\hat{\sigma}_Y}$, and the a cancels because the sample correlation is invariant to rescaling the data.

5.15 **a.** Use the standard formula to get the sample variance (here, $n = 15$):

$$\hat{\mu} = n^{-1}\sum_{i=1}^{n}X_i = 0.39,$$
$$\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2 = 0.08.$$

b. The sample mean is already unbiased.

For the variance:

$$s^2 = \frac{n}{n-1}\hat{\sigma}^2 = \frac{15}{14}0.080 = 0.086.$$

c. The mean for a $U(a,b)$ distribution is given as:

$$\hat{\mu} = \frac{a+b}{2}$$
$$0.385 = \frac{0+b}{2} \rightarrow b = 0.77.$$

d. The variance for a $U(a,b)$ distribution is given as:

$$\sigma^2 = \frac{(b-a)^2}{12}$$
$$0.086 = \frac{b^2}{12} \rightarrow b = 1.016$$

5.16 **a.** The first step is to rank order the observations:

Ranked Position	Value
1	0.26
2	0.27
3	0.28
4	0.38
5	0.43
6	0.99

From here, we anchor the first observation at 0% and the last at 100%, equally spacing the in-between values. The divisions here will be $100/n-1 =$ 20%:

Ranked Position	Value	Rank
1	0.26	0%
2	0.27	20%
3	0.28	40%
4	0.38	60%
5	0.43	80%
6	0.99	100%

The median (50%) lies exactly half way between the third and fourth ranked observations. Therefore:

$$\text{Median} = \frac{0.28 + 0.38}{2} = 0.33$$

b. This requires calculating the 25% and 75% levels.

The 25% level is 5/20 = 25% of the way between ranked observations 2 and 3. Therefore:

$$q_{25\%} = 0.75 * 0.27 + 0.25 * 0.28 = 0.2725$$

Similarly, the 75% level is 75% of the way between observations 4 and 5:

$$q_{75\%} = 0.25 * 0.38 + 0.75 * 0.43 = 0.4175$$

Thus, IQR = (0.2725, 0.4175).

Hypothesis Testing 6

Learning Objectives

After completing this reading you should be able to:

- Construct an appropriate null hypothesis and alternative hypothesis and distinguish between the two.
- Differentiate between a one-sided and a two-sided test and identify when to use each test.
- Explain the difference between Type I and Type II errors and how these relate to the size and power of a test.
- Understand how a hypothesis test and a confidence interval are related.
- Explain what the p-value of a hypothesis test measures.
- Construct and apply confidence intervals for one-sided and two-sided hypothesis tests and interpret the results of hypothesis tests with a specific confidence level.
- Identify the steps to test a hypothesis about the difference between two population means.
- Explain the problem of multiple testing and how it can lead to biased results.

As explained in Chapter 5, the true values of the population parameters are unknowable without access to the full population of data. However, observed sample data contain information about those parameters.

A *hypothesis* is a precise statement about population parameters, and the process of examining whether a hypothesis is consistent with the sample data is known as *hypothesis testing*. In frequentist inference, which is the most widely used framework for measuring the relationships between variables, hypothesis testing can be reduced to one universal question: How likely is the observed data if the hypothesis is true?

Testing a hypothesis about a population parameter starts by specifying null hypothesis and an alternative hypothesis. The *null hypothesis* is an assumption about the population parameter. The alternative hypothesis specifies the population parameter values (i.e., the *critical values*) where the null hypothesis should be rejected. These critical values are determined by:

- The distribution of the test statistic when the null hypothesis is true, and
- The size of the test, which defines the size of the rejection region and is also known as the *significance level*. This reflects the aversion to rejecting a null hypothesis that is, in fact, true.

Observed data are used to construct a test statistic, and the value of the test statistic is compared to the critical values to determine whether the null hypothesis should be rejected.

This chapter outlines the steps used to test hypotheses. It begins by examining whether the mean excess return on the stock market (i.e., the market premium) is 0. The steps used to test a hypothesis about the mean are nearly identical to those used in many econometric problems and so are widely applicable.

This chapter also examines testing the equality of the means for two sequences of random variables and explains how tests are used to validate the specification of Value-at-Risk (VaR) models.

6.1 ELEMENTS OF A HYPOTHESIS TEST

A hypothesis test has six distinct components:

1. The null hypothesis, which specifies a parameter value that is being tested and, for the purpose of the test, is assumed to be true;
2. The alternative hypothesis, which defines the range of values of the parameter where the null should be rejected;
3. The test statistic, which has a known distribution when the null is true;
4. The size of the test, which captures the size of the rejection region;
5. The critical value, which is a value that is compared to the test statistic to determine whether to reject the null hypothesis; and
6. The decision rule, which combines the test statistic and critical value to determine whether to reject the null hypothesis.

Null and Alternative Hypotheses

The first step in performing a test is to identify the null and alternative hypotheses. The null hypothesis (H_0) is a statement about the population values of one or more parameters. It is sometimes called the maintained hypothesis, because it is maintained (or assumed) that the null hypothesis is true throughout the testing procedure. Hypothesis testing relies crucially on this assumption, and many elements of a hypothesis test are based on the distribution of the sample statistic (e.g., the sample mean) when the null is true.

In most settings, the null hypothesis is often consistent with there being nothing unusual about the data. For example, when considering whether to invest in a mutual fund, the natural null hypothesis is that the fund does not generate an abnormally high return. This hypothesis can be formalized as the statement that the true average (i.e., the expected) return on the fund is equal to the true average return on its style-matched benchmark. This null hypothesis is expressed (in terms of population values) as:

$$H_0: \mu_{FUND} = \mu_{BM},$$

where μ_{FUND} is the expected return of the mutual fund (i.e., $E[r_{FUND}]$) and μ_{BM} is the expected return on the style benchmark (i.e., $E[r_{BM}]$). This null can be equivalently expressed using the difference between the means:

$$H_0: \mu_{FUND} - \mu_{BM} = 0$$

Another important type of null hypothesis occurs when testing the accuracy of an econometric model. For example, consider the testing of VaR models, which are important measures of portfolio risk. The 95% VaR of a portfolio defines a region where the daily return on the portfolio should fall 95% of the time. The area excluded from this region is always in the left tail and so captures losses to the portfolio. When testing VaR models, the null hypothesis could be that the model used to produce the VaR is correct. However, this hypothesis is not a precise statement about the value of a population parameter and so is not directly testable. Instead, the null is constructed from a key feature of a correct VaR model: The 95% VaR of a portfolio should cover actual losses on 95% of days so that the VaR is violated

on the other 5% of days. When testing a VaR model, the null hypothesis is:

$$E[HIT_t] = 5\%,$$

where a HIT_t is defined as a VaR exceedance (i.e., the loss on day t (L_t) is larger than the VaR).[1]

The next step in performing a test is to specify the alternative hypothesis (H_1), which determines the values of the population parameter(s) where the null hypothesis should be rejected. In most applications, the natural alternative is the set of values that are not covered under the null hypothesis. Returning to the example of evaluating a mutual fund's performance, the natural alternative is H_1: $\mu_{FUND} \neq \mu_{BM}$ (or equivalently, H_1: $\mu_{FUND} - \mu_{BM} \neq 0$). When testing a VaR model, the natural alternative is H_1: $E[HIT_t] \neq 5\%$. Here, rejection of the null indicates that the VaR model is not accurate: It may be either too conservative (too few violations) or too aggressive (too many violations).

One-Sided Alternative Hypotheses

In some testing problems, the alternative hypothesis might not fully complement the null. The most common example of this is called a one-sided alternative, which is used when the outcome of interest is only above or only below the value assumed by the null.

For the example of the fund manager's performance, the null is equal performance between the return on the fund and the return on the benchmark (i.e., H_0: $\mu_{FUND} = \mu_{BM}$). If the alternative was set to H_1: $\mu_{FUND} \neq \mu_{BM}$, then a rejection of the null would only indicate that the return is *different* than that of the benchmark (i.e., it could be either higher or lower than the benchmark). To focus on the case where the fund manager *outperforms* the benchmark, then a one-sided alternative is needed:

$$H_1: \mu_{FUND} > \mu_{BM}$$

or equivalently:

$$H_1: \mu_{FUND} - \mu_{BM} > 0$$

Note that since the entire area where the null hypothesis would be rejected is on one side of the distribution (rather than being split into two rejection regions), there is an increased chance of rejecting a false null hypothesis when the true parameter is in the range covered by the (one-sided) alternative.

If the one-sided test is as above yet in reality $\mu_{FUND} < \mu_{BM}$, then neither the null nor the alternative is true. However, because the values where the null is rejected are determined exclusively by the alternative hypothesis, the null should not be rejected. This highlights the fact that failing to reject the null hypothesis does not mean the null is true, but only that there is insufficient evidence against the null to be able to reject it.

One-sided alternatives are only used in specific circumstances. For example, the negative consequences of exceeding VaR limits (e.g., additional audits by regulators, loss of investor confidence, bankruptcy, and so on) may be more severe than those caused by falling short of those limits (suboptimal capital deployment). In this scenario, a one-sided test with power to detect an excessive number of large losses may be preferable to a two-sided test.

Test Statistic

The test statistic is a summary of the observed data that has a known distribution when the null hypothesis is true. Test statistics take many forms and follow a wide range of distributions. This chapter focuses on test statistics that are normally distributed, which are commonly used when the distribution of parameter estimators can be approximated using a Central Limit Theorem (CLT). This single class of test statistic can be used to test many important hypotheses, including those about means, regression coefficients, or econometric models.

Consider a test of the null hypothesis about a mean: H_0: $\mu = \mu_0$ where μ_0 is a particular numerical value being tested. When data are iid, Chapter 5 showed that the sample mean estimator is asymptotically normally distributed:

$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} N(0, \sigma^2),$$

where σ^2 is the variance of the sequence of iid random variables used to estimate the mean. When the true value of the mean (μ) is equal to the value tested by the null (μ_0), then the asymptotic distribution leads to the test statistic:

$$T = \frac{\hat{\mu} - \mu_0}{\sqrt{\hat{\sigma}^2/n}} \sim N(0,1) \tag{6.1}$$

The test statistic T is asymptotically standard normally distributed and centered at the null hypothesis. The standard normal distribution of T when the null is true is illustrated in the top-left panel of Figure 6.1. The population variance is replaced by the sample variance, because $\hat{\sigma}^2$ is a consistent estimator of the population variance σ^2 and is therefore appropriate when the sample size is sufficiently large. While this chapter focuses on testing hypotheses about the mean in most examples, T is applicable to any estimator that is asymptotically normally distributed.

[1] Formally, $HIT_t = I[L_t > VaR_t]$, where $I[\cdot]$ is an indicator function that is 1 when the argument is true (and 0 otherwise) and L_t is the portfolio loss.

Type I Error and Test Size

While hypotheses concern the true population parameters, they can only be tested using a sample. However, samples can be unrepresentative of the population and therefore produce misleading results. In an ideal world, a false (true) null would always (never) be rejected. However, in practice there is a tradeoff between avoiding a rejection of a true null and avoiding a failure to reject a false null.

Rejecting a true null hypothesis is called a Type I error.[2] The probability of committing a Type I error is known as the test size and is sometimes denoted α. The test size is chosen to reflect the willingness to mistakenly reject a true null hypothesis. The most common test size is 5% (i.e., so that there is a 1 in 20 chance that a rejected null hypothesis is actually correct). Smaller test sizes (e.g., 1% or even 0.1%) are used when it is especially important to avoid incorrectly rejecting a true null.

Critical Value and Decision Rule

The test size and the alternative are combined with the distribution of the test statistic to construct the critical value of the test. The critical value depends on the distribution of the test statistic (which is assumed normal in this chapter) and defines a range of values where the null hypothesis should be rejected in favor of the alternative. This range is known as the rejection region.

The critical value depends on both the size and the type of the alternative hypothesis (i.e., whether it is one- or two-sided). In the common case where the null is H_0: $\mu = \mu_0$ and a two-sided alternative is used (e.g., H_1: $\mu \neq \mu_0$), then the critical value defines a region in the tails with total probability α (so that the probability in each tail is $\alpha/2$).

The decision rule combines the critical value C_α (which depends on the test size α), the alternative hypothesis, and the test statistic (T) into one decision: whether to reject the null in favor of the alternative or to fail to reject the null.

When using a two-sided alternative, the decision rule is to reject the null if the absolute value of the test statistic is larger than the critical value (i.e., $|T| > C_\alpha$). When testing against a one-sided lower (upper) alternative, the decision rejects the null if $T < -C_\alpha$ ($T > C_\alpha$). When a null is rejected using a test size of α, it is said that the result is significant at the α-level.

[2] A type I error is also known as an error of the first kind.

The top left panel of Figure 6.1 shows the rejection region when $\alpha = 5\%$, which means that the shaded rejection region in each tail has a probability of 2.5%. The null is rejected in favor of the alternative hypothesis when the value of the test statistic falls into this region, whereas test statistics in the 95% probability region in the center of the distribution do not lead to rejection of the null hypothesis. Common test sizes and critical values for testing against two-sided alternatives are 10% (i.e. 5% in each tail: ± 1.645), 5% (2.5% in each tail: ± 1.96), and 1% (0.5% in each tail: ± 2.57). These critical values are obtained from the normal distribution table.

When using a one-sided alternative, the entire rejection region lies in a single tail and so is not split in two. For example, when the alternative is one-sided lower (H_1: $\mu < \mu_0$), then the critical value is chosen so that the probability in the lower tail is α. Common test sizes and critical values for one-sided lower alternatives are 10% (-1.28), 5% (-1.645), and 1% (-2.32). Test statistics less than the critical value indicate that the null should be rejected. When using a one-sided upper tailed alternative (H_0: $\mu > \mu_0$), the sign of the critical value flips and rejection occurs if the test statistic is larger than the critical value.

Example: Testing a Hypothesis about the Mean

As an example, suppose that $X = [X_1, X_2, \ldots, X_n]$ represents the annual excess return[3] on the S&P 500 from 1951 through 2018. The estimated average excess return is calculated using the realizations of X (i.e., $x = [x_1, x_2, \ldots, x_n]$) and is therefore

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

In this case, $\hat{\mu} = 7.62\%$.

To determine whether the true average excess return (i.e., $\mu = E[X_i]$) is different from 0, a hypothesis test can be constructed where:

$$H_0: \mu = 0$$

and

$$H_1: \mu \neq 0$$

Recall that the test statistic T is

$$T = \frac{\hat{\mu} - \mu_0}{\sqrt{\hat{\sigma}^2/n}} \sim N(0,1)$$

[3] The excess return is the difference between the S&P 500 and the risk-free rate. The risk-free rate is the return on an asset with zero risk, which does not exist in practice and so the return on a short-term sovereign bond is usually employed as the best available proxy. Here, the return on a one-month US government bond is used.

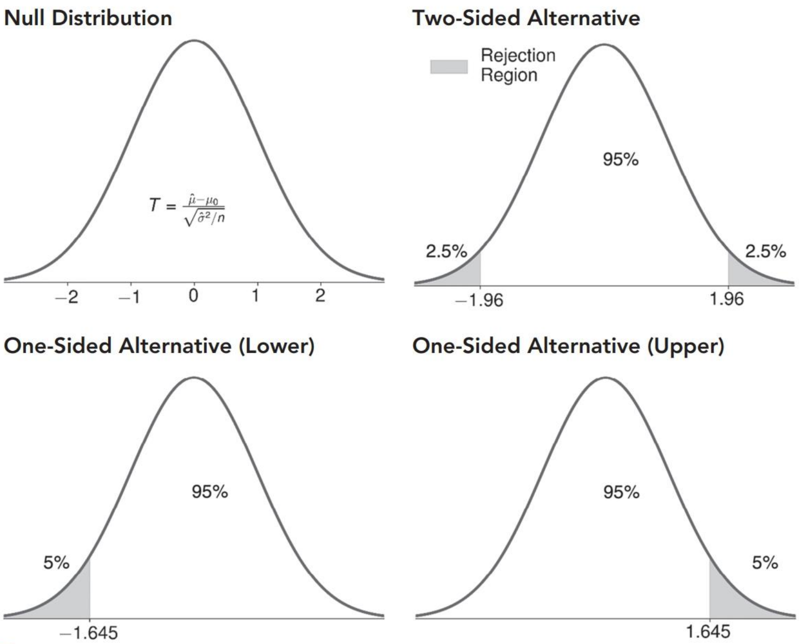


Figure 6.1 **The top-left panel shows the distribution of the test statistic, a standard normal when the null is true. The top-right panel shows the rejection region (shaded) when using a two-sided alternative and a test size of 5%. The bottom two panels show the rejection regions when using one-sided lower (left) and one-sided upper (right) alternatives also using a test size of 5%.**

THE NORMAL AND THE STUDENT'S t

This chapter uses the standard normal distribution to determine the critical values in hypothesis tests about unknown parameters. The use of the normal follows from the CLT and is applicable to averages of random variables from any distribution when some technical conditions are met.

However, there are two circumstances where the Student's t_ν should be used in place of the normal to determine the critical value(s) of a test. First, when the random variables in the average are iid normally distributed [i.e., $X_i \sim N(\mu, \sigma^2)$], then the t-test statistic has an exact Student's t_{n-1} distribution.

There is no need to use an approximation in this special case. The exact distribution requires using the unbiased estimator of the variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 = \frac{n}{n-1}\hat{\sigma}^2 \tag{6.2}$$

The test statistic for the null $H_0\colon \mu = \mu_0$ is

$$T = \frac{\hat{\mu} - \mu_0}{\sqrt{s^2/n}} \sim t_{n-1} \tag{6.3}$$

(*Continued*)

(*Continued*)

When n is small (i.e., less than 30), the Student's t has been documented to provide a better approximation to the distribution of T than the normal. This result holds even when the random variables in the average (X_i) are not normally distributed. Note that, with all else equal, the critical values from a Student's t_{n-1} are larger than those from a normal and thus using a t_{n-1} reduces the probability of rejecting the null (i.e., there would need to be more evidence against the null hypothesis before it would be rejected when the t distribution is used). In practice, the larger critical values do a better job of producing tests with the correct size, α.

Figure 6.2 compares the normal, t_5, and t_{15} densities. Note that the tails of the t are heavier than the normal, although they become closer as the degrees of freedom increase. Using critical values from the Student's t, along with the use of s^2 to estimate the variance, is the recommended method to test a hypothesis about a mean when the sample size contains 30 or fewer observations. In larger samples, the difference between the two is negligible, and common practice is to estimate the variance with $\hat{\sigma}^2$ and to use critical values from the normal distribution.

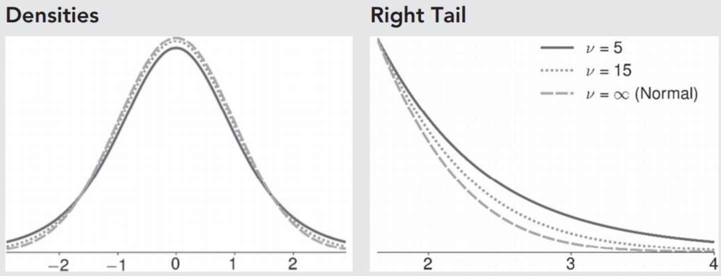


Figure 6.2 **Comparison of the densities of the normal, t_5, and t_{15}. The right panel focuses on the right tail for values above 1.645, the 5% cutoff from the normal distribution.**

In this example, the standard deviation of the excess returns is 16.5%, the sample size $n = 68$, $\hat{\mu} = 7.62\%$, and the null tests whether $\mu = 0$. Therefore, the T value for this test is

$$T = \frac{0.0762 - 0}{0.165/\sqrt{68}} = 3.8$$

Using a test size (i.e., α) of 5%, the next step is to find an interval containing 95% of the probability when the null is true. This interval is constructed by finding the critical values, which are ± 1.96 when $\alpha = 5\%$. The critical values determine the rejection region, which consists of $(-\infty, -1.96]$ and $[1.96, \infty)$. This rejection region is the shaded area in the top-right plot of Figure 6.1.

Since the test statistic value of 3.8 is in the rejection region, it is unlikely to have occurred if the null is true. Therefore, the null hypothesis (H_0: $\mu = 0$) can be rejected in favor of the alternative hypothesis (H_1: $\mu \neq 0$) and it can be concluded that the average excess return is significantly different from zero

Type II Error and Test Power

A Type II error occurs when the alternative is true (i.e., the null is wrong), but the null is not rejected. The probability of a Type II error is denoted by the Greek letter β. The power of a test, defined as $1 - \beta$, , measures the probability that a false null is rejected. In practice, β should be small so that the power of the test is high.

Table 6.1 **Decision Matrix for a Hypothesis Test Relating the Unknown Truth to the Decision Taken about the Truth of a Hypothesis. Size α and Power $1 - \beta$ Are Probabilities and so Are between 0 and 1**

		Null Hypothesis	
		True	**False**
Decision	**Fail to Reject**	Correct $(1 - \alpha)$	Type II Error (β)
	Reject	Type I Error Size: (α)	Correct Power: $(1 - \beta)$

Table 6.1 shows the decision matrix that relates the decision taken—rejecting or failing to reject the null—to the truth of the null hypothesis. Note that the values in parentheses in Table 6.1 are the probabilities of arriving at that decision.

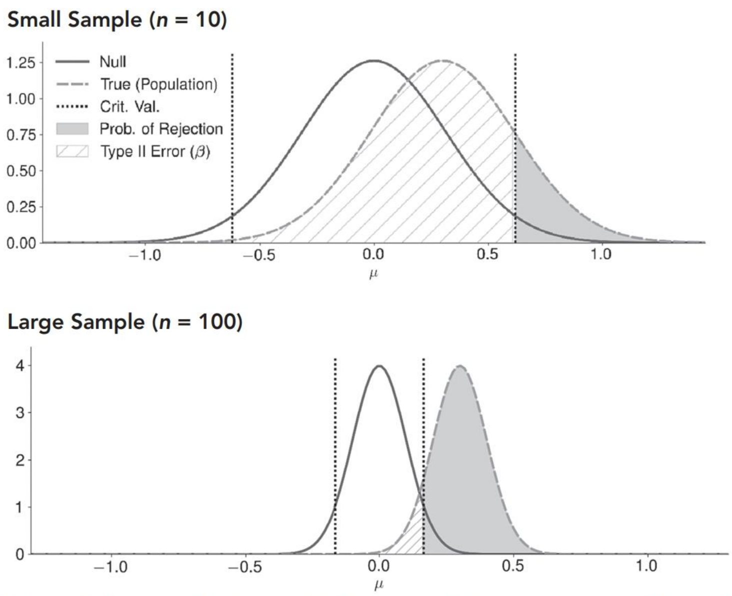


Figure 6.3 **Both panels show two distributions: the distribution of the mean when the null is true (solid), and the distribution of the sample mean that is centered at the population mean of 0.3 (dashed). The vertical lines show the critical values, which are determined by the distribution centered on the mean under consideration in the null hypothesis (0.0) assuming a one-sided test. The shaded regions show the power of the test, which is the probability that the null hypothesis is rejected in favor of the alternative. The hatched region shows the probability of making a Type II error (i.e., β).**

Whereas the size of the test can be set by the tester, the power of a test cannot be controlled in a similar fashion. This is because the power depends on:

- The sample size,
- The distance between the value assumed under the null and the true value of the parameter, and
- The size of the test.

Example: The Effect of Sample Size on Power

Figure 6.3 illustrates the core concepts of hypothesis testing in a small sample (where $n = 10$) and in a larger sample (where $n = 100$).

In both panels, the data are assumed to be generated according to:

$$X_i \overset{iid}{\sim} N(0.3, 1)$$

The sample average $\hat{\mu}$ is the average of n iid normal random variables and so is also normally distributed with $E[\hat{\mu}] = 0.3$ and $V[\hat{\mu}] = 1/n$. The null hypothesis is H_0: $\mu = 0$ and the alternative is H_1: $\mu \neq 0$.

Both panels in Figure 6.3 shows two distributions. The first is the distribution of the sample mean if the null is true (shown with the solid line). It is centered at 0 and has a standard deviation of $1/\sqrt{10}$. The second is the distribution of the sample mean (shown with the dashed line). It is centered at the true value of $\hat{\mu} = 0.3$ and has the same standard deviation.

The dashed lines show the critical values for a test with a size of 5%, which are:

- ± 0.61, when $n = 10$, and
- ± 0.196, when $n = 100$

for the top and bottom panels, respectively. These numbers are obtained by taking the critical value from the normal distribution at the 5% level (1.96) and dividing it by $\sqrt{n}$, which is $\sqrt{10}$ in the first case and $\sqrt{100}$ in the second. Note that because the sampling distributions are being compared to the mean, the diagrams plot the mean on the *x*-axis rather than the test statistic *T*. Instead of scaling the sample mean with its standard error, as would be done to compute *T*, the critical value is multiplied by the standard error (which in this case is $1/\sqrt{n}$ since $\sigma^2 = 1$).

When the sample mean falls outside of the set of critical values, the null is rejected in favor of the alternative. The probability that the sample mean falls into the rejection region is illustrated by the grey region outside of the critical values. This area is the power of the test (i.e., the chance that the null is rejected when the data is generated under the alternative). Meanwhile, the probability of failing to reject the null (i.e., β) is indicated by the hatched area.

Increasing the sample size has an obvious effect—both distributions are narrower. These narrower distributions increase the probability that the null is correctly rejected and so increase the power of the test. In even larger samples (e.g., $n = 1000$), the two distributions would have virtually no overlap so that the null is rejected with an exceedingly high probability (>99.999%). This pattern shows that the power (i.e., the probability that the false null is rejected) increases with *n*.

Example: The Effect of Test Size and Population Mean on Power

Figure 6.4 shows how the power of a hypothesis test is affected by the true (population) mean and the test size. In this case, the null hypothesis is H_0: $\mu = 0$ and the test uses sizes of 1%, 5%, and 10%. The curves show the power for the sample size $n = 50$.

Note how test size influences power, because larger sizes (e.g., 10% rather than 5%) use smaller critical values (i.e., closer to zero) and have larger rejection regions than smaller tests sizes (e.g., 1%). Smaller critical values increase the probability that the sample mean falls into the rejection region, and so increase the test power.

Note also that changing the sample test size has an important side effect: Large test sizes increase the probability of committing a Type I error and rejecting a true null hypothesis. The choice of the test size therefore affects both the probability of making a Type I error and that of making a Type II error. In other words, small sizes reduce the frequency of Type I errors, while large sizes reduce the frequency of Type II errors. The test size should thus be chosen to reflect the costs and benefits of committing either a Type I or Type II error.

Note that the power is equal to α (i.e., the test size) when the true value of μ is close to 0. This makes sense, because the null should be rejected with probability α even when true.

As the true μ moves away from 0, however, the power increases monotonically. This is because it is easier to tell that the null is false when the data are generated by a population μ far from the value assumed by the null (since the null is "more wrong" in such cases). In other words, large effects are easier to detect than small effects.

Confidence Intervals

A confidence interval is a range of parameters that complements the rejection region. For example, a 95% confidence interval contains the set of parameter values where the null hypothesis cannot be rejected when using a 5% test. More generally, a $1 - \alpha$ confidence interval contains the values that cannot be rejected when using a test size of α.

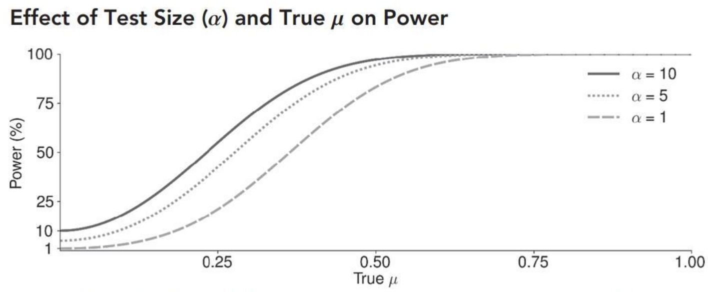


Figure 6.4 **The graph shows the effect of changing the test size on power using $n = 50$ observations.**

The bounds of a confidence interval depend on the type of alternative used in the test. When H_0: $\mu = \mu_0$ is tested against the two-sided alternative (e.g., H_1: $\mu \neq \mu_0$), then the $1 - \alpha$ confidence interval is

$$[\hat{\mu} - C_\alpha \times \hat{\sigma}/\sqrt{n}, \hat{\mu} + C_\alpha \times \hat{\sigma}/\sqrt{n}], \tag{6.4}$$

where C_α is the critical value for a two-sided test (e.g., if $\alpha = 5\%$ then $C_\alpha = 1.96$).

Returning to the example of estimating the mean excess return on the S&P 500, the 95% confidence interval is

$$\left[7.62 - 1.96 \times \frac{16.5}{\sqrt{68}}, 7.62 + 1.96 \times \frac{16.5}{\sqrt{68}}\right] = [3.69, 11.54]$$

This confidence interval indicates that any value of μ_0 under the null between 3.69 and 11.54 cannot be rejected against a two-sided alternative. Other values (e.g., 0) are outside of this confidence interval, and so the null H_0: $\mu = 0$ is rejected. Thus, confidence intervals provide a different approach for testing hypotheses compared with constructing test statistics as described previously. It is possible to show that one approach is just a rearrangement of the other and both always provide the same conclusion.

Confidence intervals can also be constructed for one-sided alternatives. The one-sided $1 - \alpha$ confidence interval is asymmetric and is either:

$$(-\infty, \hat{\mu} + C_\alpha \times \hat{\sigma}/\sqrt{n}]$$

or

$$[\hat{\mu} - C_\alpha \times \hat{\sigma}/\sqrt{n}, \infty),$$

where C_α is the critical value for a one-sided test with a size of α. Returning to the excess return on the stock market, the one-sided confidence interval for testing the null H_0: $\mu = 0$ against the alternative H_1: $\mu > 0$ is

$$\left[7.62 - 1.645 \times \frac{16.5}{\sqrt{68}}, \infty\right) = [4.32, \infty)$$

The critical value is reduced from 1.96 to 1.645 due to the directionality of the test. In the two-sided test, the rejection region was shared between the left and right tails. In the one-sided test, the rejection region is only in the left tail. The null that $\mu = 0$ is outside of this one-sided confidence interval as well and so is rejected.

One-sided confidence intervals highlight a surprising issue when using one-sided tests: The sign of the mean estimate $\hat{\mu}$ affects the decision to reject. For example, when testing H_0: $\mu = 0$ against the one-sided alternative H_1: $\mu > 0$, *any* negative estimate of $\hat{\mu}$ does not lead to rejection of the null hypothesis.

While confidence intervals can be interpreted as the set of null values that would not be rejected using a test size of α, they can also be correctly interpreted using a repeated sampling argument. Suppose that many independent samples are generated and that the $1 - \alpha$ confidence interval is constructed for each sample. It is the case that $1 - \alpha$ (e.g., 95%) of these confidence intervals will contain the true parameter when the number of repeated samples is large. This second interpretation is more of a thought experiment in practice because it is usually not possible to create multiple independent samples of data from financial markets.

The p-value of a Test

A hypothesis test can also be summarized by its p-value, which measures the probability of observing a test statistic that is more extreme than the one computed from the observed data when the null hypothesis is true.

A p-value combines the test statistic, distribution of the test statistic, and the critical values into a single number that is always between 0 and 1. This value can always be used to determine whether a null hypothesis should be rejected: If the p-value is less than the size of the test, then the null is rejected.

The p-value of a test statistic is equivalently defined as the smallest test size (α) where the null is rejected. Any test size larger than the p-value leads to rejection, whereas using a test size smaller than the p-value fails to reject the null. Roughly speaking, the p-value represents the probability of drawing an incorrect conclusion when the null hypothesis is rejected.

The p-value depends on whether the alternative is one- or two-sided. When the alternative is two-sided (H_1: $\mu \neq \mu_0$), the p-value of a test statistic value T is

$$2(1 - \Phi(|T|)), \tag{6.5}$$

where $\Phi(z)$ is the CDF of a standard normal evaluated at z.

In this expression, $|T|$ is used to ensure that the probability in the right-tail is always measured, irrespective of whether T is positive or negative. Therefore, $\Phi(|T|)$ computes the probability $\Pr(Z \leq |T|)$ and $1 - \Phi(|T|)$ computes the area to the right of $|T|$ under the normal density ($\Pr(Z > |T|)$, where Z is a standard normal random variable. Finally, this area is doubled when using a two-sided test, because the test statistic T may lie in either tail.

In the example testing whether the excess return on the stock market is 0 (H_0: $\mu = 0$), the test statistic value is:

$$T = 3.80$$

Therefore, the p-value can be calculated as:

$$\begin{aligned} 2(1 - \Phi(|3.80|)) &= 2(1 - 0.9993) \\ &= 2(.00007) \\ &= 0.00014 \end{aligned}$$

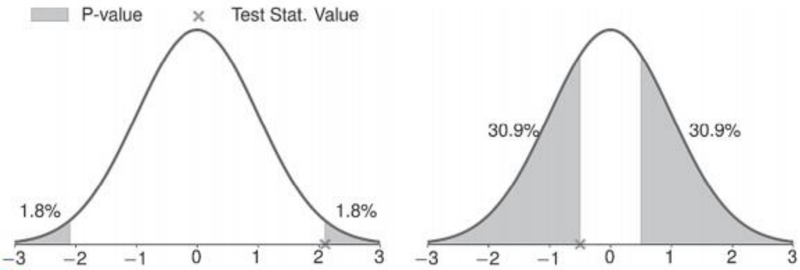


Figure 6.5 **Graphical representation of p-values. The two panels show the area measured by a p-value when using a two-sided alternative. The p-value measures the area above $|T|$ and below $-|T|$ when using the test statistic T.**

PRACTICAL VERSUS STATISTICAL SIGNIFICANCE

Hypothesis tests establish whether a finding is statistically significant (i.e., whether the null hypothesis is rejected using an appropriate test size). The hypothesis test does not, however, indicate whether a finding has practical significance. In sufficiently large samples, standard hypothesis tests have the property that the null hypothesis is always rejected whenever null is false and the alternative is true. This happens because any discrepancy, no matter how small, must produce a test statistic that grows arbitrarily large as n increases, because the test statistic can be expressed

$$T = \sqrt{n}\frac{\hat{\mu} - \mu_0}{\hat{\sigma}}$$

In large samples, $\hat{\mu}$ converges to the true value of μ so that when the null is false:

$$\mu - \mu_0 \neq 0$$

As n increases, the test statistic takes the form:

$$\sqrt{n}\delta,$$

where δ is a non-zero value and so always falls into the rejection region. As a result, very large samples produce statistically significant results even when the alternative is extremely close to the null.

Practical significance applies a second check when assessing whether a result is noteworthy. Practical significance examines the magnitude of the estimated effect (i.e., the difference between μ and μ_0), not magnitude of the test statistic or its p-value. It complements statistical significance, because for a result to be actionable it must be both statistically and practically significant.

For example, suppose a new portfolio strategy reduces portfolio variance by 0.1% per year while not affecting the performance. Given a long enough historical sample, this difference is always statistically significant. However, if the portfolio is a high-risk strategy with a volatility of 40% per year, then this minor reduction in volatility is unlikely to alter the attractiveness of the portfolio to new investors. Here the result is statistically, but not practically, significant and thus is not actionable.

The null hypothesis is rejected because 0.014% is less than the test size of 5%. And since the p-value is so small, it can be concluded that it is extremely unlikely a test statistic as large as this would have occurred by chance alone.

In addition, the one-sided p-value for testing the null against the alternative H_1: $\mu > 0$ is $1 - \Phi(3.80) = 0.007\%$, which is half of the two-sided p-value.

Figure 6.5 shows the area measured by the p-value when the alternative is two-sided. The test statistic and the tail probabilities in the left panel are 2.1 and 1.8%, respectively, making the p-value equal to 3.6%. In this case, the null hypothesis would be rejected using a test size of 5% (but not 1%). The p-value of 3.6% indicates that there is a 3.6% probability of observing a $|T| \geq 2.1$ the null hypothesis is true.

In the right panel, the test statistic is −0.5 and the total probability in both tails (i.e., the p-value) is 61.8%. This test would not reject the null using any standard test size. The large p-value here indicates that when the null is true, three out of five samples would produce an absolute test statistic as large as 0.5. (i.e., it is quite likely that a test statistic of this value would be generated by chance alone).

The formula for a one-sided p-value depends on the direction of the alternative hypothesis. When using a one-sided lower-tailed test (e.g., H_1: $\mu < \mu_0$), the p-value is $\Phi(T)$ (i.e., the area

THE RISKS OF REPEATED TESTING

Multiple testing (i.e., testing multiple hypotheses on the same data) is a pervasive problem that affects fields as diverse as finance, psychology, biology, and medical research. Reusing data can produce spurious results and findings that do not hold up to rigorous scrutiny.

The fundamental problem with multiple testing is that the test size (i.e., the probability that a true null is rejected) is only applicable for a single test. However, repeated testing creates test sizes that are much larger than the assumed size of α and therefore increases the probability of a Type I error.

For example, suppose there are n trading strategies that are being compared to a buy-and-hold strategy. These strategies can be tested to determine which (if any) provide a better return than the buy-and-hold approach. The test would involve n nulls of the form:

$$H_0: \mu_i = \mu_{BaH,}$$

where μ_i is the return on strategy i and μ_{BaH} is the return from the buy-and-hold strategy.

If the test statistics are independent, then there is a 40.1% (=100% − (95%)^{10}) chance that one will appear statistically significant at the 5% level when $n = 10$. This probability increases to 99.4% when $n = 100$. In others words: If you keep testing strategies on the same data, you are bound to eventually find one that appears to work. However, this finding would be spurious and it would be very likely to fail if you then retested the strategy using different data.

That being said, there are ways to control for multiple testing. The leading methods are the Bonferroni (or Holm-Bonferroni) correction, which corrects the decision rule of a testing procedure when using multiple tests. It is simple to implement because it only makes use of standard statistics and their p-values. As a result, Holm-Bonferroni can be very conservative and have low power.

Other methods can involve controlling the False Discovery Rate (FDR) or limiting the Familywise Error Rate (FWER). FDR and FWER are more complex approaches and use sophisticated, data-intensive methods to overcome this loss of power.

in the left tail). When using a one-sided upper-tailed test (e.g., $H_1: \mu > \mu_0$), the p-value is $1 - \Phi(T)$ (i.e., the area in the right tail). The p-values for one-sided alternatives do not use the absolute value because the sign of the test statistic matters for determining statistical significance.

6.2 TESTING THE EQUALITY OF TWO MEANS

Testing whether the means of two series are equal is a common problem. Consider the iid bivariate random variable $W_i = [X_i, Y_i]$. Since W_i is a bivariate variable, observations on X_i and Y_i occur in pairs and hence the number of observations of each (i.e., n) are equal. The component random variables X_i and Y_i are each iid and may be contemporaneously correlated (i.e., $\text{Corr}[X_i, Y_i] \neq 0$).

Now consider a test of the null hypothesis $H_0: \mu_X = \mu_Y$ (i.e., that the component random variables X_i and Y_i have equal means). To implement a test of this null, construct a new random variable:

$$Z_i = X_i - Y_i$$

If the null hypothesis is true, then:

$$E[Z_i] = E[X_i] - E[Y_i] = \mu_x - \mu_y = 0$$

This is a standard hypothesis test of $H_0: \mu_Z = 0$ against the alternative that $H_1: \mu_z \neq 0$.

The test statistic is constructed as:

$$T = \frac{\hat{\mu}_Z}{\sqrt{\hat{\sigma}_Z^2/n}}$$

and its value can be compared to a standard normal. This method automatically accounts for any correlation between X_i and Y_i because:

$$V[Z_i] = V[X_i] + V[Y_i] - 2\,\text{Cov}[X_i, Y_i]$$

Note that because $\hat{\mu}_Z = \hat{\mu}_X - \hat{\mu}_Y$ and $\hat{\sigma}_Z^2 = \hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}$, this test statistic can be equivalently expressed as:

$$T = \frac{\hat{\mu}_X - \hat{\mu}_Y}{\sqrt{\dfrac{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}{n}}}$$

This expression for T shows that the correlation between the two series affects the test statistic. If the two series are positively correlated, then the covariance between X_i and Y_i reduces the variance of the difference. This occurs because both random variables tend to have errors in the same direction, and so the difference between $X_i - Y_i$ eliminates some of the common error.

As an example, consider two distributions that measure the average rainfall in City X and City Y, respectively. Their means, variances, sample size and correlation are:

$$\mu_X = 10\text{cm};\ \mu_Y = 14\text{cm};\ \sigma_X^2 = 4,\ \sigma_Y^2 = 6,\ n_X = n_Y = 12,\ \text{Corr}[X_i, Y_i] = 0.30$$

For the hypothesis to not be rejected at the 95% level, the maximum difference between μ_X and μ_Y would be:

$$T = \frac{|D|}{\sqrt{\frac{4 + 6 - 2 * 0.3 * \sqrt{4}\sqrt{6}}{12}}} \leq 1.96 => |D| \leq 2.6cm$$

However, because:

$$|\mu_X - \mu_Y| = |-4| > 2.6$$

The null is therefore rejected when $\alpha = 0.05$.

Furthermore, note that the value of the test statistic is

$$T = \frac{\hat{\mu}_X - \hat{\mu}_Y}{\sqrt{\frac{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}{n}}}$$

$$= \frac{10 - 14}{\sqrt{\frac{4 + 6 - 2 * 0.3 * \sqrt{4}\sqrt{6}}{12}}}$$

$$= -5.21$$

Therefore, the p-value is <0.01% and the null would be rejected even at a 99.99% confidence level!

Finally, a special version of this test statistic is available when X_i and Y_i are both iid and mutually independent. When X_i and Y_i are independent, they are not paired as a bivariate random variable and so the number of sample points for X_i (n_X) and Y_i (n_Y) may differ. The test statistic for testing that the means are equal (i.e., H_0: $\mu_X = \mu_Y$) is[4]

$$T = \frac{\hat{\mu}_X - \hat{\mu}_Y}{\sqrt{\frac{\hat{\sigma}_X^2}{n_X} + \frac{\hat{\sigma}_Y^2}{n_Y}}}$$

6.3 SUMMARY

This chapter introduces hypothesis testing. Implementing a hypothesis test involves specifying the null and alternative hypotheses. The null hypothesis is an assumption about the population parameters being tested. The alternative hypothesis defines the values where the null hypothesis should be rejected. Most applications use a two-sided alternative that is the natural complement to the null. The hypothesis test is implemented with a test statistic that has a known distribution when the null is true.

The size of the test is chosen to control the probability of a Type I error, which is the rejection of a true null hypothesis. The test size should reflect the cost of being wrong when the null is true. The distribution of the test statistic and the test size are combined with a critical value, which defines the rejection region. Test statistics in the rejection region lead to rejection of the null. If the test statistic is not in the rejection region, the null is not rejected.

Confidence intervals and p-values are alternative but equivalent methods to express the results of a testing procedure. Confidence intervals contain a set of values where the null hypothesis should not be rejected. The p-value is the probability of observing a test statistic as large as the one computed using the data if the null is true. This means that a null should be rejected if the p-value is smaller than the test size of α.

Most of the focus of this chapter is on testing a null hypothesis about the mean (H_0: $\mu = \mu_0$) using the CLT. More generally, this testing framework is applicable to any parameter that has an asymptotic distribution that can be approximated using a CLT.

[4] This test statistic also has a standard normal distribution when the null is true.

The following questions are intended to help candidates understand the material. They are _not_ actual FRM exam questions.

QUESTIONS

Short Concept Questions

6.1 What role does the alternative hypothesis play when testing a hypothesis?

6.2 When is a one-sided alternative useful?

6.3 What is the size of a hypothesis test?

6.4 What are the trade-offs when choosing the size of a test?

6.5 What is the power of a hypothesis test?

6.6 What is the critical value in a hypothesis test?

6.7 What does the p-value of a hypothesis measure?

6.8 How can a confidence interval be used when testing a hypothesis?

6.9 What does the VaR of a portfolio measure?

6.10 What are three methods to evaluate a VaR model of a portfolio?

Practice Questions

6.11 Suppose you wish to test whether the default rate of bonds that are classified as investment grade by S&P is the same as the default rate on bonds classified as investment grade by Fitch. What are the null and alternative hypotheses? What data would you need to test the null hypothesis?

6.12 Using the Excel function NORM.S.DIST or a normal probability table, what are the critical values when testing the null hypothesis, H_0: $\mu = \mu_0$, against

- **a.** a one-sided lower alternative using a size of 10%?
- **b.** a one-sided upper alternative using a size of 20%?
- **c.** a two-sided alternative using a size of 2%?
- **d.** a two-sided alternative using a size of 0.1%?

6.13 Find the p-value for the following *t*-test statistics of the null hypothesis, H_0: $\mu = \mu_0$.

- **a.** Statistic: 1.45, Alternative: Two-sided
- **b.** Statistic: 1.45, Alternative: One-sided upper
- **c.** Statistic: 1.45, Alternative: One-sided lower
- **d.** Statistic: −2.3, Alternative: One-sided upper
- **e.** Statistic: 2.7, Alternative: Two-sided

6.14 If you are given a 99% confidence interval for the mean return on the Nasdaq 100 of [2.32%, 12.78%], what is the sample mean and standard error? If this confidence interval is based on 37 years of annual data, assumed to be iid, what is the sample standard deviation?

6.15 You collect 50 years of annual data on equity and bond returns. The estimated mean equity return is 7.3% per year, and the sample mean bond return is 2.7% per year. The sample standard deviations are 18.4% and 5.3%, respectively. The correlation between the two-return series is −60%. Are the expected returns on these two assets statistically significantly different from each other? Does your answer change if the correlation is 0?

6.16 If a *p*-VaR model is well specified, *HITs* should be iid Bernoulli$(1 - p)$. What is the probability of observing two *HITs* in a row? Can you think of how this could be used to perform a test that the model is correct?

6.17 A data management group wants to test the null hypothesis that observed data are N(0,1) distributed by evaluating the mean of a set of random draws. However, the actual underlying data are distributed as N(1, 2.25).

- **a.** If the sample size is 10, what is the probability of a Type II error and the power of the test? Assume a 90% confidence level on a two-sided test.
- **b.** How many data points would need to be taken to reduce the probability of a Type II error to less than 1%?

6.18 Suppose that for a linear regression, an estimate of the slope is stated as having a one-sided p-value of 0.04. What does this mean?

The following questions are intended to help candidates understand the material. They are ***not*** actual FRM exam questions.

ANSWERS

Short Concept Questions

6.1 The alternative specifies the range of values where the null should be rejected. In most tests, the alternative is the natural complement to the null, although this does not have to be the case.

6.2 A one-sided alternative is helpful when there is some guidance from economic or financial theory or from prior intuition about the value of the parameter when the null is false. A leading example occurs when testing whether a risk premium is 0 against a one-sided alternative that the premium is positive.

6.3 The size is the probability of wrongly rejecting the null when it is in fact true, which is also the size of the rejection region. This is equivalently the probability of a Type I error.

6.4 A small test size reduces the probability of a Type I error, that is, rejecting the null when it is true. It also lowers the power of the test, that is the chance of rejecting a false null because smaller sizes correspond to larger critical values.

6.5 The power of a test the probability of rejecting a false null hypothesis when the alternative is true. It is one minus the probability of a Type II error.

6.6 The critical value determines the rejection region for a test and is found from the tables of statistical distributions. Test statistics more extreme (i.e., larger in absolute value) than the critical value indicate the null should be rejected in favor of the alternative.

6.7 The p-value measures the probability that the observed data would have been generated if the null is true. It is the size (marginal significance level) where the test would just reject the null in favor of the alternative.

6.8 If the value of the parameter under the null is outside of the confidence interval, then the null should be rejected in favor of the alternative.

6.9 The VaR is a measure of the magnitude of the loss that the portfolio will lose with some specified probability (e.g., 5%) over some fixed horizon (e.g., one day or one week). The *p*-VaR is formally defined as the value where:

$$\Pr(L > VaR) = 1 - p, \tag{6.6}$$

where *L* is the loss of the portfolio of the selected time horizon and $1 - p$ is the probability that a large loss occurs.

6.10 First, the exact distribution that computes the exact distribution of the number of VaR violations (*HITs*) over some time period. Second, the asymptotic method that examines the average number of violations and uses the CLT to constrict the asymptotic distribution. Finally, the likelihood ratio that exploits the Bernoulli distribution that underlies the 0-1 *HIT* variables.

Solved Problems

6.11 The null is that the IG (investment grade) default rate is the same for S&P rated firms as it is for Fitch-rated firms. If S_i are default indicators for S&P rated firms, and F_i are default indicators for Fitch-rated firms, then the null is H_0: $\mu_S = \mu_F$, which states that the mean values are the same. The null can be equivalently expressed as H_0: $\mu_S - \mu_F = 0$. The alternative is H_1: $\mu_S \neq \mu_F$ or equivalently H_1: $\mu_S - \mu_F \neq 0$. The data required to test this hypothesis would be binary random variables where 1 indicates that an IG bond defaulted and 0 if it did not within a fixed time frame (e.g., a quarter).

6.12 **a.** The lower-tailed critical value is -1.28.

b. The one-sided upper-critical value is 0.84.

c. A two-sided alternative with a size of 2% uses the same critical value as a one-sided upper test, which is 2.32. Each tail has probability 1% when using the critical value.

d. This size corresponds to the same critical value as a 0.05% upper-tailed test, which is 3.29.

6.13 **a.** In a two-sided test, the p-value is the area under the normal curve for values greater than 1.45 or less than -1.45. This is twice the area less than -1.45, or $2 \times 0.074 = 0.148$.

b. In a one-sided upper test, we need the area above 1.45. This is half the previous answer, or 0.074.

c. In a one-sided lower test, we need the area under the curve for values less than 1.45. This is just $1 - 0.074 = 0.926$.

d. Here we need to probability under the normal curve for values above -2.3, which is 1 minus the area less than -2.3, or $1 - 0.01 = 0.99$.

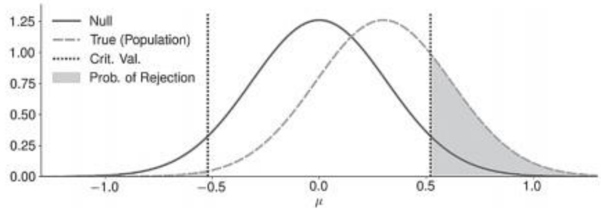
e. Because it is a two-sided alternative, we need that area for values larger than 2.7 or less than −2.7, which is twice the area less than −2.7. This value is $2 \times 0.0034 = 0.0068$.

6.14 The mean is the midpoint of a symmetric confidence interval (the usual type), and so is 7.55%. The 99% CI is constructed as $[\hat{\mu} - c \times \hat{\sigma}, \hat{\mu} + c \times \hat{\sigma}]$ and so $c \times \hat{\sigma} = 12.78\% - 7.55\% = 5.23\%$. The critical value for a 99% CI corresponds to the point where there is 0.5% in each tail, or 2.57, and so $\hat{\sigma} = \frac{5.23\%}{2.57} = 2.03\%$.

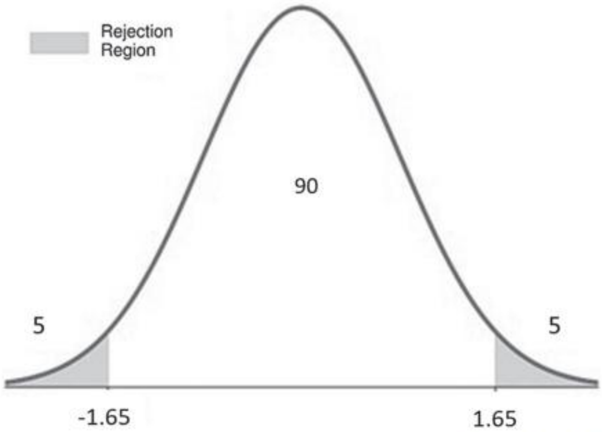
6.15 The null hypothesis is H_0: $\mu_E = \mu_B$. The alternative is H_1: $\mu_E \neq \mu_B$. The test statistic is based on the difference of the average returns, $\hat{\delta} = 7.3\% - 2.7\% = 4.6\%$. The estimator of the variance of the difference is $\hat{\sigma}_E^2 + \hat{\sigma}_B^2 - 2\hat{\sigma}_{BE}$, which is $0.184^2 + 0.053^2 - 2 \times -0.6 \times 0.184 \times 0.053 = 0.0483$. The test statistic is $\frac{\hat{\delta}}{\sqrt{\frac{0.043}{50}}} = 1.47$. The critical value for a two-sides test is ± 1.96 using a size of 5%. The null is not rejected. If the correlation was 0, then the variance estimate would be 0.0366, and the test statistic would be 1.69. The null would still not be rejected if the size was 5%, although if the test size was 10%, then the critical value would be ± 1.645 and the correlation would matter.

6.16 The probability of a *HIT* should be $1 - p$ if the model is correct. They should also be independent, and so the probability of observing two *HIT*s in a row should be $(1 - p)^2$. The can be formulated as the null that H_0: $E[HIT_i \times HIT_{i+1}] = (1 - p)^2$ and tested against the alternative H_1: $E[HIT_i \times HIT_{i+1}] \neq (1 - p)^2$. This can be implemented as a simple test of a mean by defining the random variable $X_i = HIT_i \times HIT_{i+1}$ and then testing the null H_0: $\mu_X = (1 - p)^2$ using a standard test of a mean.

6.17 **a.** When the null hypothesis is false, the probability of a Type II error is equal to the probability that the hypothesis fails to be rejected, as per the diagram in the chapter:



Further recall that for a 90% confidence level on a N(0,1) distribution, the cut-off points are +/−1.65.



Now, if there are 10 data points taken from an N(0,1) then the standard deviation is reduced:

$$\hat{\sigma}_{H_0} = \frac{1}{\sqrt{10}} = 0.316$$

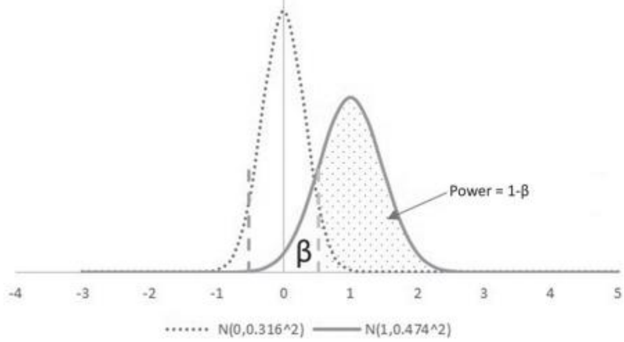
Therefore, the cut-off points are:

$$\pm 1.65 \times 0.316 = \pm 0.522$$

In actuality, the true distribution is N(1,2.25), so $\sigma = \sqrt{2.25} = 1.5$. For a sample size of 10, the expected sample standard deviation is:

$$\hat{\sigma}_{sample} = \frac{1.5}{\sqrt{10}} = 0.474$$

Calculating the equivalent distance of +/− 1.65 in this distribution compared to a standard N(0,1) yields:



$$\text{left} = \frac{-0.522 - 1}{0.474} = -3.21$$

and

$$\text{right} = \frac{+0.522 - 1}{0.474} = -1.00$$

The probability of being on the left-hand side is practically zero. For the right:

$\Pr(> right) = 1 - \Phi(-1.00) = 1 - 15.9\% = 84.1\%$.

So the total probability of a Type II error is 1 – the probability of being in the two tails is:

$$\Pr(Non - Rejection|\ H_o \text{ is false}) = 1 - [\Pr(< left) + \Pr(> right)] \approx 1 - 84.1\% = 15.9\%$$

Therefore, the power of the test is 84.1%.

b. The requirement is to have:

$$1 - [\Pr(< left) + \Pr(> right)] = 1\%$$

Clearly, as *n* increases from 10, the probability of being in the left-hand tail will only decrease from already being close to zero.

Therefore, the requirement becomes:

$$1 - \Pr(> \text{right}) = 0.01$$

This occurs at a Z-score of (using the Excel function NORMSINV) −2.32.

Accordingly, the following equations need to be solved:

$$1.65 * \hat{\sigma}_{H_0} = \frac{1.65}{\sqrt{n}} = K \text{ and } \frac{+K - 1}{\left(\frac{1.5}{\sqrt{n}}\right)} = -2.32$$

Plugging in *K* yields:

$$\frac{\left(\frac{1.65}{\sqrt{n}}\right) - 1}{\left(\frac{1.5}{\sqrt{n}}\right)} = \frac{1.65 - \sqrt{n}}{1.5} = -2.32 \geq \sqrt{n} = 5.13$$

$$\geq n = 26.3$$

And because partial observations are not allowed, $n = 27$.

6.18 The p-value tells us the "probability of observing a test statistic that is more extreme than the one computed from the observed data when the null hypothesis is true." In other words, if the estimate of the slope is true, then in a randomized trial the probability of getting a larger new slope estimate is 4%.

Assuming a normal distribution, this will correspond to the value of x that solves the following equation:

$$1 - \Phi(x) = 0.04$$

Using the excel function NORMSINV gives $x = 1.75$. So a one-sided p-value of 4% would correspond to a test statistic of 1.75 for a normally distributed random variable. We could also state that there is a 4% probability that a test statistic as large as this or larger could have occurred by chance alone.

7 Linear Regression

Learning Objectives

After completing this reading you should be able to:

- Describe the models that can be estimated using linear regression and differentiate them from those which cannot.
- Interpret the results of an ordinary least squares (OLS) regression with a single explanatory variable.
- Describe the key assumptions of OLS parameter estimation.
- Characterize the properties of OLS estimators and their sampling distributions.
- Construct, apply, and interpret hypothesis tests and confidence intervals for a single regression coefficient in a regression.
- Explain the steps needed to perform a hypothesis test in a linear regression.
- Describe the relationship among a t-statistic, its p-value, and a confidence interval.

Linear regression is a widely applied statistical tool for modeling the relationship between random variables. It has many appealing features (e.g., closed-form estimators, interpretable parameters, and a flexible specification) and can be adapted to a wide variety of problems. This chapter develops the bivariate linear regression model, which relates a dependent variable to a single explanatory variable. Models with multiple explanatory variables are examined later in this book.

The chapter begins by examining the specifications that can (and cannot) be modeled in the linear regression framework. Regression is surprisingly flexible and can (using carefully constructed explanatory variables) describe a wide variety of relationships. Dummy variables and interactions are two key tools used when modeling data. These are widely used to build flexible models where parameter values change depending on the value of another variable.

This chapter presents the Ordinary Least Squares (OLS) estimators, which have a simple moment-like structure and depend on the mean, variance, and covariance of the data. While these estimators are derived without assumptions about the process that generated the data, assumptions are needed to interpret the model. Five of these are used to establish key statistical properties of linear regression estimators. When satisfied, the parameter estimators are asymptotically normal, and standard inference can be used to test hypotheses.

Finally, this chapter presents three key applications of linear regression in finance: measuring asset exposure to risk factors, hedging, and evaluating fund manager performance.

7.1 LINEAR REGRESSION

Regression analysis is the most widely used method to measure, model, and test relationships between random variables. It is widely used in finance to measure the sensitivity of a portfolio to common risk factors, estimate optimal hedge ratios for managing specific risks, and to measure fund manager performance, along with many other applications.

Regression models all examine the relationship between a dependent variable (Y) and one or more explanatory variables (X). This chapter focuses on the most widely applied form of regression modeling: linear regression with a single explanatory variable, sometimes known as bivariate regression or simple regression.

Linear Regression Parameters

Linear regression assumes a linear relationship between an explanatory variable X and a dependent variable Y so that:

$$Y = \alpha + \beta X + \epsilon, \tag{7.1}$$

VARIABLE NAMING CONVENTIONS

Linear regression is used widely across finance, economics, other social sciences, engineering, and the natural sciences. While users within a discipline often use the same nomenclature, there are differences across disciplines. The three variables in a regression are frequently referred to using a variety of names. These are listed in the table below.

Explained Variable	Explanatory Variable	Shock
Left-Hand-Side Variable	Right-Hand-Side Variable	Innovation
Dependent Variable	Independent Variable	Noise
Regressand	Regressor	Error
		Disturbance

where:

- β, commonly called the slope or the regression coefficient, measures the sensitivity of Y to changes in X;
- α, commonly called the intercept, is a constant; and
- ϵ, commonly called the shock/innovation/error/disturbance, represents a component in Y that cannot be explained by X. This shock is assumed to have mean 0 so that:

$$E[Y] = E[\alpha + \beta X + \epsilon] = \alpha + \beta E[X]$$

The presence of the shock is also why statistical analysis is required. If the shock were not present, then Y and X would have a perfect linear relationship and, when plotted, the data points for (X, Y) would all lie on a perfect straight line so that the value of the two parameters could be exactly determined using only basic algebra. Note that the variables Y and X are treated asymmetrically since causality runs from X and Y and not the other way around. Contrast this framework with correlation, where the two variables are treated symmetrically.

An obvious interpretation of $\hat{\alpha}$ is the value of Y when X is zero. However, this interpretation is not meangingful if X cannot plausibly be zero.

For example, consider a model that examines how the average maturity of corporate bond offerings (Y) depends on firm market capitalization (X). Because firm size is always positive, the observed data always lie to the right of the y-axis and interpreting $\hat{\alpha}$ as the average maturity of debt offered by firms with a market capitalization of USD 0 is not meaningful. Moreover, it may be the case that the prediction at 0 is not even feasible (e.g., a negative value of $\hat{\alpha}$ would

be incorrect in the maturity example because maturity is also always positive).

When the explanatory variable cannot be 0, then $\hat{\alpha}$ is interpreted as the value that ensures that $\overline{Y}$ lies on the fitted regression line at $\overline{X}$ (i.e., $\overline{Y} = \hat{\alpha} + \hat{\beta}\overline{X}$). This interpretation is always meaningful because x_i can always lie on either side of $\overline{X}$.

Linearity

Explanatory variables can be continuous, discrete, or functions of one or more variables (e.g., $X_3 = X_1 \times X_2$). However, all linear regressions must satisfy three essential restrictions.

1. First, the relationship between Y and the explanatory variables $X_1, X_2, \ldots, X_k$ must be linear in the unknown coefficients (i.e., α and β). This means that the term on the right-hand-side of the model must have a single unknown coefficient multiplied by a single explanatory variable.
2. Second, the error must be additive. This restriction excludes some models where the variance of the error depends on observed data (e.g., $Y = \alpha + \beta X + \gamma X\epsilon$).
3. Finally, all explanatory variables must be observable. This limitation precludes directly applying linear regression with missing data.[1]

For example, consider

$$Y = \alpha + \beta X^{\gamma} + \epsilon,$$

where γ is an unknown parameter that measures the shape of the nonlinear relationship. The parameters of this model cannot be estimated using the methodology of linear regression because βX^{γ} contains two unknown parameters and γ does not enter multiplicatively (and therefore it violates the first restriction).

However, the use of *multiple regression* (i.e., multiple explanatory variables) can allow some nonlinear relationships to be modeled using linear regression. The multiple regression model includes k explanatory variables so that:

$$Y = \alpha + \beta_1 X_1 + \ldots + \beta_k X_k + \epsilon \qquad (7.2)$$

Building models with multiple explanatory variables allows the effect of an explanatory variable to be measured while controlling for other variables known to be related to Y. For example, when assessing the performance of a hedge fund manager, it is common to use between seven and 11 explanatory variables that capture various sources of risk.

[1] There are many approaches to imputing missing values in linear models that then allow linear regression to be applied to the combined dataset including the imputed values.

Note that the k explanatory variables are not assumed to be independent, and so one variable can be defined as a known nonlinear function of another.[2] Transforming variables enables linear regression to explain a wide range of relationships between the dependent and the explanatory variables. For example, it is possible to model the relationship between Y and the pair X and X^2 in a linear regression by setting $X_1 = X$ and $X_2 = X^2$, so that:

$$\begin{aligned} Y &= \alpha + \beta_1 X + \beta_2 X^2 + \epsilon \\ &= \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon \end{aligned}$$

Here, the second explanatory variable is a function of the first, and the model is not linear in X, because a small increase in X by ΔX increases Y by:[3]

$$\Delta Y = (\beta_1 + 2\beta_2 X)\Delta X$$

Note that rate of change of Y depends on both the change ΔX and the level of X.

Many smooth nonlinear functions can be accurately approximated using polynomials (i.e., functions of X, X^2, X^3, etc.) and so these relationships can be modeled using linear regression by including powers of explanatory variables. These models are considered linear in the parameters (e.g., α and β) but nonlinear in the variables (e.g., X).

Transformations

Some model specifications that do not naively satisfy the three requirements for linearity can be transformed so that they do. For example, suppose that Y is an always positive random variable, and that Y, X, and ϵ are related through:

$$Y = \alpha X^{\beta}\epsilon,$$

where ϵ is a positive-valued shock. This specification is not compatible with the requirements of linear regression, because X is raised to an unknown coefficient β and the error ϵ is not additive. However, taking the natural logarithm of both sides of the equation transforms the relationship so that:

$$\begin{aligned} \ln Y &= \ln \alpha + \beta \ln X + \ln \epsilon \\ \widetilde{Y} &= \widetilde{\alpha} + \beta\widetilde{X} + \widetilde{\epsilon} \end{aligned}$$

[2] While the k explanatory variables are not assumed to be independent, they are assumed not to be perfectly correlated. This structure allows for nonlinear transformation of a variable to be included in a regression model, but not for linear transformation because $\text{Corr}[X_i, a + bX_i] = 1$ if $b > 0$ and -1 if $b < 0$. Further details in the assumptions made with using multiple explanatory variables are presented in Chapter 10.

[3] This is the result of taking the partial derivative of Y with respect to X.

This transformed model satisfies the three requirements of linear regression. When interpreting the slope of a transformed relationship, note that the coefficient β measures the effect of a change in *the transformation of X on the transformation of Y.*

For example, if the model is:

$$\ln Y_i = \alpha + \beta \ln X + \epsilon_i$$

then β is the change in ln Y for a small change in ln X.

Computing the direct effect of X on Y requires taking the derivative of the functions transforming these variables. In the case of ln Y, the derivative is:

$$\partial Y/Y$$

which can be directly interpreted as a percentage change. When both X and Y have been replaced by their logarithms (called a log-log model), then β measures the percentage change in Y for a 1% change in X. In other words, the coefficient in a log-log model is directly interpretable as the elasticity of Y with respect to X.

However, if only the dependent variable is replaced by its logarithm (e.g., $\ln Y_i = \alpha + \beta X_i + \epsilon$), then β measures the percentage change in Y for a one-unit change in X.

Dummy Variables

An important class of explanatory variable is known as a dummy. A dummy random variable is almost invariably binary and only takes the value 0 or 1. Dummies are used to encode qualitative information (e.g., a firm's sector or a bond's country of origin). Dummies can also be used to capture seasonal patterns in the relationships between variables.

A dummy is usually specified to take the value 1 when the observation has the quality and 0 if it does not. For example, when encoding sectors, the transportation sector dummy is 1 for a firm whose primary business is transportation (e.g., a commercial airline or a bus operator) and 0 for firms outside the industry. Dummies are also commonly constructed as binary transformations of other random variables (e.g., a market direction dummy that encodes the return on the market as 1 if negative and 0 if positive).

Dummies can be added to models to change the intercept (α) or the slope (β) when the dummy variable takes the value 1. For example, an empirically relevant extension of CAPM allows for the sensitivity of a firm's return to depend on the direction of the market return. If we define X to be the excess market return and D to be a dummy variable that is 1 if the excess market return is negative (i.e., a loss), then this asymmetric CAPM specifies the relationship as

$$\begin{aligned} Y &= \alpha + \beta_1 X + \beta_2(X \times D) + \epsilon \\ &= \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon \end{aligned}$$

This model is a linear regression because the two variables ($X_1 = X$ and $X_2 = X \times D$) are both observable and each is related to the dependent variable through a single coefficient. The second explanatory variable ($X \times D$) is called an interaction dummy variable, because it is produced by interacting a binary dummy variable with another variable.

The dummy interaction makes it possible to describe two models with a single equation:

- $Y = \alpha + \beta_1 X + \epsilon$, when $D = 0$; and
- $Y = \alpha + (\beta_1 + \beta_2)X + \epsilon$, when $D = 1$.

Even though each segment is linear in X (i.e., it is piecewise linear), the slopes of the two lines differ when $\beta_2 \neq 0$ and thus the overall relationship between Y and X is nonlinear. The dummy variable here has the effect of altering the slope of the regression equation and so it is known as a slope dummy variable. If instead the dummy was included additively in the model, it would allow the intercept to shift and so would be known as an intercept dummy:

$$Y = \alpha + \alpha_1 D + \beta X + \epsilon$$

It is also possible to include both intercept and slope dummy variables in the same equation.

7.2 ORDINARY LEAST SQUARES

Consider a bivariate regression model that includes a single explanatory variable:

$$Y = \alpha + \beta X + \epsilon$$

This model has three parameters: α (i.e., the intercept), β (i.e., the slope), and σ^2 (i.e., the variance of ϵ).

Suppose that there are n observations of Y and X. All pairs (x_i, y_i) are assumed to have the same linear relationship so that:

$$y_i = \alpha + \beta y_i + \epsilon_i$$

These parameters can be estimated using Ordinary Least Squares (OLS) estimators, which are derived by minimizing the sum of squared deviations of the points to the fitted line. The deviations are squared so that positive and negative values do not cancel each other out.

In this case, the squared deviations being minimized are between the realizations of the dependent variable Y and their expected values given the respective realizations of X:

$$\hat{\alpha}, \hat{\beta} = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2, \qquad (7.3)$$

where arg min denotes argument of the minimum.[4]

In turn, these estimators minimize the residual sum of squares, which is defined as:

$$\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = \sum_{i=1}^{n}\hat{\epsilon}_i^2$$

In other words, the estimators (i.e., $\hat{\alpha}$ and $\hat{\beta}$) are the intercept and slope of a line that best fits the data because it minimizes the squared deviations between the line $\hat{\alpha} + \hat{\beta}x_i$ and the realizations of the dependent variable y_i. Equation (7.3) is solved by differentiating the expression with respect to α and β, setting the derivative to zero, and rearranging the resulting equation.

The solution to the minimization problem is

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{X})^2}, \tag{7.4}$$

where $\overline{Y}$ and $\overline{X}$ are the averages of y_i and x_i, respectively.

Note that the estimator $\hat{\beta}$ is only sensible if $\sum_{i=1}^{n}(x_i - \overline{X})^2 > 0$, which requires that x_i has some variation around its mean. If x_i only takes one value, then all points lie along a vertical line and the slope is infinite. Violations of this assumption are simple to identify in a dataset and so are easily avoided.

The estimators $\hat{\alpha}$ and $\hat{\beta}$ are then used to construct the fitted values:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$$

and the model residuals (i.e., the differences between the actual and fitted values for each data point *i*)

$$\hat{\epsilon}_i = y_i - \hat{y}_i = y_i - \hat{\alpha} - \hat{\beta}x_i$$

Finally, the variance of the shocks is estimated by:

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\epsilon}_i^2 \tag{7.5}$$

Note that the variance estimator divides by $n - 2$ to account for the two estimated parameters in the model, so that s^2 is an unbiased estimator of σ^2 (i.e., $E[s^2] = \sigma^2$).

The residuals are always mean 0 (i.e., $\sum_{i=1}^{n}\hat{\epsilon}_i = 0$) and uncorrelated with X_i (i.e., $\sum_{i=1}^{n}(X_i - \overline{X})\hat{\epsilon}_i = 0$). These two properties are consequences of minimizing the sum of squared errors. There is a technical distinction between the model residuals ($\hat{\epsilon}_i$) and the population disturbances (ϵ_i), which are never observed.

The estimator for the slope can be equivalently expressed as the ratio of the sample covariance between *X* and *Y* to the variance of *X*, because we can multiply both the numerator and denominator by $\frac{1}{n}$ to get:

$$\hat{\beta} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_{XX}} \tag{7.6}$$

This ratio can also be rewritten in terms of the correlation and the standard deviations so that:

$$\hat{\beta} = \hat{\rho}_{XY}\frac{\hat{\sigma}_y}{\hat{\sigma}_x} \tag{7.7}$$

The sample correlation $\hat{\rho}_{xy}$ uniquely determines the direction of the line relating *Y* and *X*. It also plays a role in determining the magnitude of the slope. In addition, the sample standard deviations $\hat{\sigma}_y$ and $\hat{\sigma}_x$ scale the correlation in the final determination of the regression coefficient $\hat{\beta}$. This relationship highlights a useful property of OLS: $\rho = 0$ if and only if $\beta = 0$. Thus, linear regression is a convenient method to formally test whether data are correlated.

Figure 7.1 illustrates the components of a fitted model that regresses the mining sector portfolio returns on the market return for 20-year period between 1999 and 2018. The solid line is the estimated regression line, which minimizes the sum of the squared *vertical* distances between the observed data and the fitted line.

The mean annual excess returns of the mining sector (*Y*) and the market (*X*) are $\overline{y} = 15.4\%$ and $\overline{x} = 7.77\%$, respectively. The covariance matrix for the two annual returns is

$$\begin{bmatrix} \hat{\sigma}_Y^2 & \hat{\sigma}_{XY} \\ \hat{\sigma}_{XY} & \hat{\sigma}_X^2 \end{bmatrix} = \begin{bmatrix} 1770 & 512.3 \\ 512.3 & 337.3 \end{bmatrix}$$

The estimate of the slope can be calculated using Equation (7.6):

$$\hat{\beta} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_{XX}} = \frac{512.3}{337.3} = 1.52$$

The intercept depends on the slope and the two estimated means, and is calculated as

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x} = 15.4\% - 1.52 \times 7.77\% = 3.60\%$$

The estimated slope shows that the return on the mining sector portfolio increases by an average of 1.52% when the market return increases by 1%. Meanwhile, the estimated intercept indicates that the portfolio returns an average of 3.607% when the market return is zero.

The square marker in Figure 7.1 shows the fitted value of the portfolio return in 2003. Note that the observed mining portfolio market return is far above the fitted value, and so the residual (marked by the brace) is positive. It can be concluded that the model fits this data point particularly poorly.

[4] For example, $\arg\min_x f(x)$ is the value of *X* the produces that lowest value of *f*(*x*).

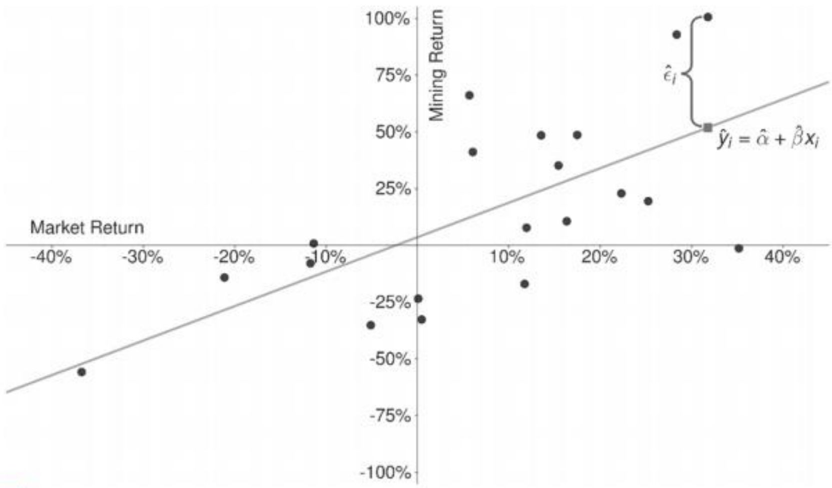


Figure 7.1 Plot of the annual returns on a sector portfolio that tracks the performance of mining companies and the returns on the US equity market between 1999 and 2018. The diagonal line is the fitted regression line where the portfolio return is regressed on the market return. The square point is the fitted value for a particular data point when using the market return in 2003 of 31.8%. The fund's return was 100.5%. The brace shows the residual, which is the error between the observed value and the regression line.

7.3 PROPERTIES OF OLS PARAMETER ESTIMATORS

The derivation of OLS estimators requires only one easy-to-verify assumption—that the variance of X is positive. However, five additional assumptions are needed to establish conditions that ensure the OLS estimators are interpretable and have desirable statistical properties.

Shocks Are Mean Zero

Shocks are mean zero conditional on X, so that $E[\epsilon|X] = 0$. This property is known as *mean independence* and it requires that X has no information about the *location* of ϵ.

This assumption states that:

$$E[\epsilon g(X)] = 0,$$

where $g(X)$ is any well-defined function of X, including both the identity function ($g(X) = X$) and the constant function ($g(X) = 1$). This assumption implies that $\text{Corr}[\epsilon, X] = 0$, so that the innovations are uncorrelated with the explanatory variables. This assumption also implies that the unconditional mean of the shocks is zero ($E[\epsilon] = 0$).

This assumption is not directly testable, as shocks are not observable and the estimated residuals (i.e., $\hat{\epsilon}_i$) are always exactly uncorrelated with the observations of the explanatory variable (i.e., x_i). Determining whether this assumption is reasonable requires a careful examination of the data generating process for (Y, X). Examples of data generating processes where this assumption is violated include the following.

- *Sample selection bias or survivorship bias.* Sample selection bias occurs when some observations are not recorded due to missing values of y_i. For example, when modeling housing price changes using completed transactions, it is important to remember that homeowners often avoid selling when they have negative equity (i.e., when the value of the home is less than the outstanding mortgage). This effect removes homes with prices that have fallen the most from the sample and creates sample selection bias. Similarly, when studying firm performance, the firms in operation are "winners" in the sense that they have performed well enough to continue to do business. Firms that are less successful are more likely to delist, and any model estimated using only the surviving firms is not representative of the behavior of all firms. Survivorship bias is commonly addressed using carefully constructed databases that report the final return for all firms, including those that delist.

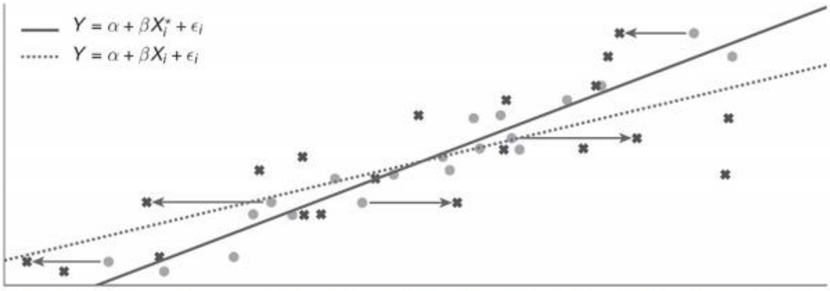


Figure 7.2 **An illustration of attenuation bias. The explanatory variable X_i^* (dots) is not observed and the model is instead estimated using noisy measurements, $X_i = X_i^* + \eta_i$ (crosses), where η_i is independent measurement noise. The arrows show the random shift from the true value of the explanatory variable to the value used in the regression.**

Both these biases imply that the sample would no longer be representative of the underlying population.

- *Simultaneity bias.* The standard regression framework assumes that causality runs only from X to Y (i.e., that changes in X lead to changes in Y and not the other way around). But when X and Y are simultaneously determined, then modeling Y as a function of X is not meaningful because X is also a function of Y. The classic example of simultaneity bias is the relationship between quantities transacted and the transaction price. Changing the transaction price affects the quantity traded, which in turn affects the price.
- *Omitted variables.* The model should not exclude variables that are important determinants of Y. Omitting these creates coefficients on the included variables that are biased and may indicate a relationship when there is, in fact, none.
- *Attenuation bias.* As stated above, it is assumed that the explanatory variables are non-stochastic. When they are measured with error, the estimated slopes are smaller in magnitude than the true slopes. This is because measurement error attenuates the relationship between Y and X, leading to inconsistent parameter estimates. Figure 7.2 shows the effect of attenuation bias using simulated data. The true explanatory variable is X_i^* (marked with dots) and is not observed. Instead, a noisy measure (marked with crosses) of it is used

$$X_i = X_i^* + \eta_i,$$

where η_i is independent of X_i^* and the shock in the model. These measurement errors randomly shift the explanatory variables and produce a line with a slope that is always flatter than the true relationship. Measurement error is common in finance when using balance-sheet variables reported on quarterly statements. Attenuation bias has been called the *iron law of econometrics*, which suggests that estimated coefficients systematically understate the true strength of relationships due to the pervasiveness of measurement error.

Data Are Realizations from iid Random Variables

When using financial and economic data, explanatory variables are not fixed values that can be set or altered. Rather, an explanatory variable is typically produced from the same process that produces the dependent variable.

For example, the explanatory variable in CAPM is the market return and the dependent variable is the return on a portfolio. Both the return on the portfolio and the market return are the result of market participants incorporating new information into their expectations of future prices. Thus, CAPM models the mean of the excess return on the portfolio *conditioned* on the excess return on the market.

Formally, it is assumed that the pairs (x_i, y_i) are iid draws from their joint distribution. This assumption allows x_i and y_i to be simultaneously generated. Importantly, the iid assumption affects the uncertainty of the OLS parameters estimators because it rules out correlation across observations. In some applications of OLS (e.g., sequentially produced time series data), the iid assumption is violated. Note that OLS can still be used in situations where (X, Y) are not iid, although the method used to compute standard errors of the estimators must be modified.

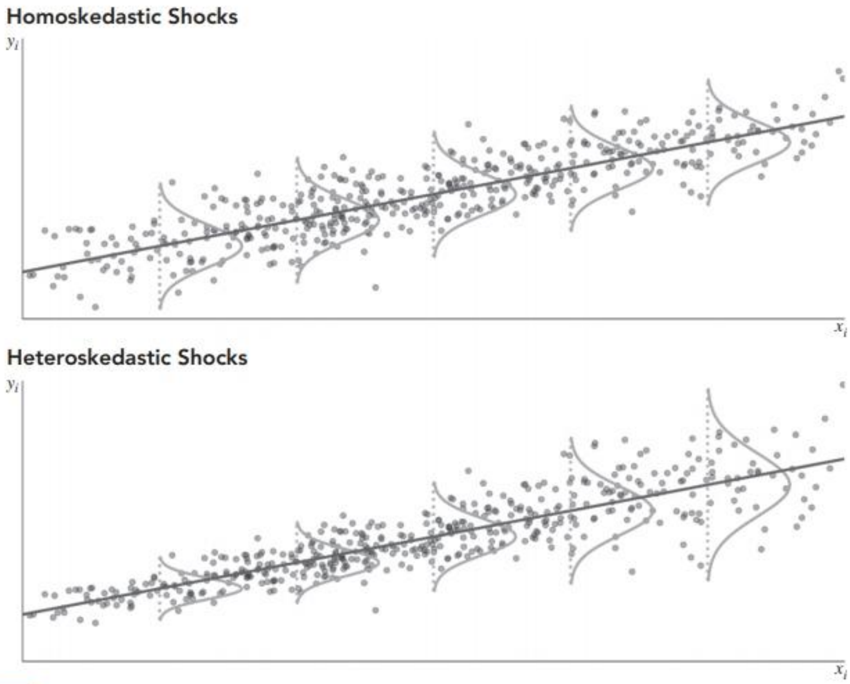


Figure 7.3 **The densities along the regression line illustrate the distribution of the shocks for five values of X_i. In the top panel, these distributions are identical and the shocks in the regression are homoskedastic. The bottom panel shows heteroskedastic data where the variance of the shock is an increasing function of the value of X_i.**

Variance of X

The variance of X is strictly greater than 0 (i.e., $\sigma_X^2 > 0$). This assumption is required when estimating the parameters and ensures that $\hat{\beta}$ is well defined.

Constant Variance of Shocks

The variance of the shocks is finite and does not vary with X_i so that:

$$V[\epsilon_i|X] = \sigma^2 < \infty \quad (7.8)$$

This assumption is known as homoskedasticity and requires that the variance of all shocks is the same.

The top panel of Figure 7.3 illustrates this assumption with simulated data. The density of the shocks is shown for five values of x_i. Note that this distribution does not vary and so the variance of the shocks is constant. However, the bottom panel of Figure 7.3 illustrates a violation of this assumption. The density of the errors is changing systematically with x_i, and the variance of the shocks (which is reflected by the width of the interval) is increasing in the explanatory variable. When the variance of the shocks is not constant, the shocks are considered to be *heteroskedastic*.

No Outliers

The probability of large outliers in X should be small.[5] OLS minimizes the sum of squared errors and therefore is sensitive to large deviations in either the shocks or the explanatory variables. Outliers lead to large increases in the sum of squared errors, and parameters estimated using OLS in the presence of outliers may differ substantially from the true parameters.

[5] Formally, we assume that the random variables X_i have finite fourth moments. This technical condition cannot be directly verified using a sample of data, and so this technical detail is of limited value.

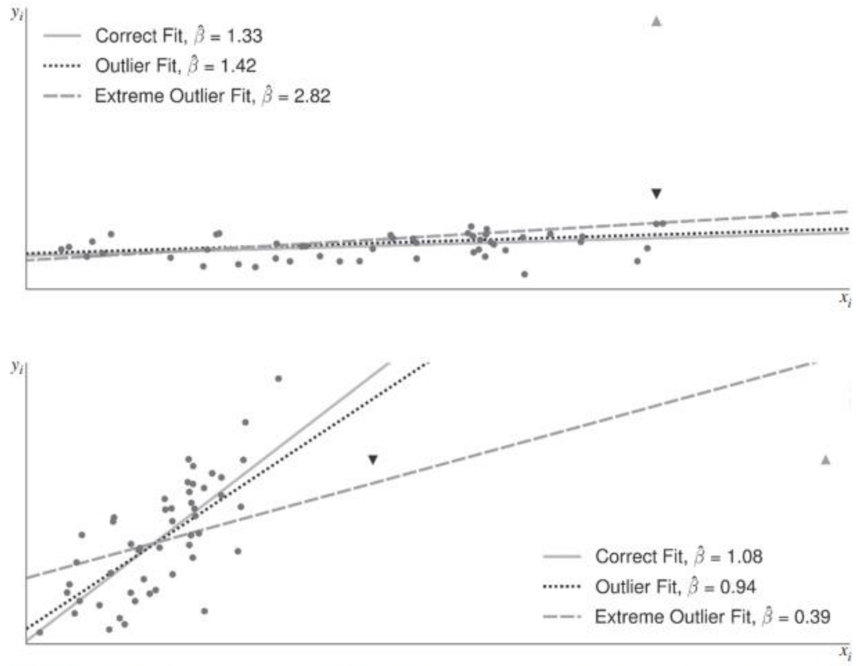


Figure 7.4 **The top panel shows the effect of a single outlier in ϵ_i on the estimated slope coefficient of a moderate outlier and an extreme outlier. The bottom panel shows the effect on the slope of an outlier and an extreme outlier in the explanatory variable.**

Figure 7.4 shows examples of outliers in ϵ (top panel) and X (bottom panel). The single outlier in the error has the potential to substantially alter the fitted line. Here, the extreme outlier leads to a doubling of the slope despite the 49 other points all occurring close to the true line. The bottom panel shows that outlier in X can also produce a large change in the estimated slope. In this case, the outlier reduces the estimate of the slope by more than 50%.

The simplest method to detect and address outliers is to visually examine data for extreme observations by plotting both the explanatory variables and the fitted residuals.

Implications of OLS Assumptions

These assumptions come with two meaningful implications. First, they imply that the estimators are unbiased so that:

$$E[\hat{\alpha}] = \alpha \text{ and } E[\hat{\beta}] = \beta$$

Unbiased parameter estimates are (on average) equal to the true parameters. This is a finite sample property and so holds for any number of observations n. In large samples, the estimators are consistent so that $\hat{\alpha} \xrightarrow{p} \alpha$ and $\hat{\beta} \xrightarrow{p} \beta$ as $n \rightarrow \infty$. This property ensures that the estimated parameters are very close to the true values when n is large.

The second meaningful implication is that the two estimators are jointly normally distributed.

The asymptotic distribution of the estimated slope is:

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N\left(0, \frac{\sigma^2}{\sigma_X^2}\right), \tag{7.9}$$

where $V[X] = \sigma_X^2$ is the variance of X.

This CLT strengthens the consistency result and allows for hypothesis testing. Note that the variance of the slope estimator depends on two moments: the variance of the shocks (i.e., σ^2) and the variance of the explanatory variables (i.e., σ_X^2). Furthermore, the variance of $\hat{\beta}$ increases with σ^2. This is not surprising,

because accurately estimating the slope is more difficult when the data are noisy.

On the other hand, the variance of $\hat{\beta}$ is decreasing in σ_X^2. This reflects the value of having widely dispersed explanatory variables; all things equal, it is easier to identify the slope when the observations of X are spread out.

The asymptotic distribution of the intercept is

$$\sqrt{n}(\hat{\alpha} - \alpha) \xrightarrow{d} N\left(0, \frac{\sigma^2(\mu_X^2 + \sigma_X^2)}{\sigma_X^2}\right) \quad (7.10)$$

The estimation error in $\hat{\alpha}$ also depends on the variance of the residuals and the variance of X. In addition, it depends on μ_X^2 (i.e., the squared mean of X). If X has mean 0 (i.e., $\mu_X = 0$), then the asymptotic variance in Equation (7.10) simplifies to σ^2 and $\hat{\alpha}$ simply estimates the mean of Y.

In practice, the CLT is used as an approximation so that $\hat{\beta}$ is treated as a normal random variable that is centered at the true slope β:

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{n\sigma_X^2}\right) \quad (7.11)$$

The effect of the sample size n is clear in this approximation: the variance of $\hat{\beta}$ decreases as the sample size increases.

ESTIMATING THE STANDARD ERROR OF THE OLS PARAMETERS

Because the expression in Equation (7.11) depends on the unknown values σ^2 and σ_X^2, it cannot be used to test hypotheses. Rather, these values are replaced by the estimator of the variance of the shocks (i.e., s^2) and the large sample estimator of the variance of the observations of X:

$$\hat{\sigma}_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2 \quad (7.12)$$

Note that $n\hat{\sigma}_X^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2$ and thus the standard form of the estimator of the variance of $\hat{\beta}$ is

$$\widehat{V[\beta]} = \frac{s^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{s^2}{n\hat{\sigma}_X^2} \quad (7.13)$$

The estimated standard error of $\hat{\beta}$ is then:

$$\widehat{\text{s.e.}(\beta)} = \frac{s}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}} = \frac{1}{\sqrt{n}}\frac{s}{\hat{\sigma}_X} \quad (7.14)$$

The estimator for the variance of $\hat{\alpha}$ can be constructed by replacing population quantities with their sample analogues, so that:

$$\widehat{V[\alpha]} = \frac{s^2(n^{-1}\sum X_i^2)}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{s^2(\hat{\mu}_X^2 + \hat{\sigma}_X^2)}{n\hat{\sigma}_X^2}, \quad (7.15)$$

where the numerator uses a simplification derived from the identity $E[X_i]^2 = \mu_X^2 + \sigma_X^2$.

7.4 INFERENCE AND HYPOTHESIS TESTING

When the assumptions in Section 7.3 are satisfied, the estimators of α and β are normally distributed in large samples.[6] Therefore, testing a hypothesis about a regression parameter is identical to testing a hypothesis about the mean of a random variable (or any other estimator that follows a CLT). Tests are implemented using a t-test, which measures the normalized difference between the estimated parameter and the value specified by the null hypothesis. When testing the null $H_0{:}\beta = \beta_0$, the test statistic is

$$T = \frac{\hat{\beta} - \beta_0}{\text{s.e.}(\hat{\beta})} = \frac{\hat{\beta} - \beta_0}{\frac{s}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}} \quad (7.16)$$

When the null hypothesis is true, then the test statistic has an asymptotic normal distribution. The test is implemented by comparing the test statistic value to the critical values from a standard normal distribution. When the alternative is two-sided (e.g., $H_1{:}\beta \neq \beta_0$), then the null is rejected when:

$$|t| > C_s,$$

where C_s is the critical value from the normal distribution for a test with size s. For example, when using a 5% test, the critical value is chosen so that the probability in each tail is 2.5%, and so $C_s = 1.96$.

The test statistic can also be transformed into a p-value, which measures the probability of observing a test statistic as large as the one observed if the null is true. The p-value is computed by measuring the area in both tails (for a two-sided test) or in the single tail (for a one-sided test) that is beyond the test statistic. It is computed by first measuring the area in the right tail beyond the test statistic, $1 - \Phi(|T|)$, and then doubling this to account for both tails so that:

$$\text{p-value} = 2(1 - \Phi(|T|)), \quad (7.17)$$

where $\Phi(T)$ is the CDF of a standard normal. A p-value less than 5% indicates that the null is rejected using a test size of 5%.

The asymptotic distribution can also be used to construct $1 - c$ confidence intervals using the quantiles of a standard normal

[6] The parameters are jointly asymptotically (bivariate) normally distributed so that:

$$\sqrt{n}\begin{bmatrix}\hat{\alpha} - \alpha \\ \hat{\beta} - \beta\end{bmatrix} \xrightarrow{d} N\left(\begin{bmatrix}0 \\ 0\end{bmatrix}, \begin{bmatrix}\frac{\sigma^2(\mu_X^2 + \sigma_X^2)}{\sigma_X^2} & -\frac{\sigma^2\mu_X}{\sigma_X^2} \\ -\frac{\sigma^2\mu_X}{\sigma_X^2} & \frac{\sigma^2}{\sigma_X^2}\end{bmatrix}\right)$$

This joint distribution allows the testing of joint hypotheses that involve both α and β. In most applications, β is the parameter of interest. The parameter α is included in the regression model only to prevent model misspecification, which would affect the estimated value of β.

TESTING THE SLOPE IN A LINEAR REGRESSION MODEL

The asymptotic variance of the slope depends on the variance of the explanatory variable and the variance of the shocks. These two values can be estimated using only the elements of the covariance matrix between the dependent and the explanatory variables. It can be shown that:

$$\sum_i \hat{\epsilon}_i^2 = n(\hat{\sigma}_Y^2 - \hat{\beta}^2\hat{\sigma}_X^2)$$

and so the variance of the innovations can be estimated using:

$$s^2 = \frac{n}{n-2}\hat{\sigma}_Y^2(1 - \hat{\rho}_{XY}^2),$$

where $\hat{\rho}_{XY}$ is the sample correlation between the dependent and the explanatory variables. Using the estimates in the mining sector example in Section 7.2, the estimated shock variance is:

$$s^2 = \frac{20}{18} \times 1770\,(1 - .663^2) = 991$$

The asymptotic variance of $\hat{\beta}$ is then:

$$\frac{s^2}{\hat{\sigma}_X^2} = \frac{991}{337.3} = 2.93$$

The standard error of $\hat{\beta}$ is:

$$\sqrt{2.93/n} = \sqrt{2.93/20} = 0.383$$

This value can be used to construct the test statistic of $\hat{\beta}$, which tests the null $H_0{:}\beta = 0$:

$$\frac{\hat{\beta}}{0.383} = \frac{1.52}{0.383} = 3.97$$

The absolute value of the test statistic is larger than the critical value for a test with a size of 5% (i.e., 1.96 s), and so the null is rejected. The estimated coefficient and its standard error can also be combined to construct a two-sided 95% confidence interval:

$$\left[\hat{\beta} - 1.96 \times \widehat{\text{s.e.}(\beta)},\ \hat{\beta} + 1.96 \times \widehat{\text{s.e.}(\beta)}\right]$$
$$= [1.52 - 1.96 \times 0.383,\ 1.52 + 1.96 \times 0.383]$$
$$= [0.769,\ 2.27]$$

The confidence interval does not include 0, which reconfirms the previous finding that the null is rejected when using a 5% test.

Finally, the p-value of the slope is:

$$2(1 - \Phi(|3.76|)) = 2(1 - 0.99996) = 0.0001$$

This value is less than 5%. Thus, the p-value provides another method to test the null and shows it is highly improbable that the true population value of the slope parameter is zero.

distribution.[7] Recall that a $1 - c$ confidence interval contains the set of null hypothesis values that are *not* rejected when using a test size of c. For example, the 90% confidence interval for β is

$$\left[\hat{\beta} - 1.645 \times \widehat{\text{s.e.}(\beta)},\ \hat{\beta} + 1.645 \times \widehat{\text{s.e.}(\beta)}\right], \qquad (7.18)$$

where $\Pr(-1.645 \leq Z \leq 1.645) = 90\%$ when Z is a standard normal random variable.

It is common to report the *t*-statistic of the regression coefficient, which is the value of the test statistic for the specific null hypothesis that the parameter is 0.

For example, the *t*-statistic of β has the null $H_0{:}\beta = 0$ and alternative $H_1{:}\beta \neq 0$. Values larger than ± 1.96 (or 2, using the rule-of-thumb value) indicate that the parameter is statistically different from 0. Many software packages also report the p-value of the *t*-statistic and a 95% confidence interval for the estimated parameter. These can equivalently be used to determine whether a parameter is statistically different from 0 (i.e., the null is rejected if 0 is not in the 95% confidence interval or if the p-value is less than 5%).

[7] In Chapter 6, we called this a $1 - \alpha$ confidence interval, where α is the size of the test. c has the same interpretation as α in the previous chapter and is used here to avoid ambiguity with the intercept α in the linear regression model that has been introduced in this chapter using the conventional notation.

7.5 APPLICATION: CAPM

The CAPM relates the excess return on a portfolio to the excess return on the market, so that:

$$R_p - R_f = \alpha + \beta(R_{mi} - R_{fi}) + \epsilon_i, \qquad (7.19)$$

where R_p is the return on a portfolio, R_m is the market return, and R_{fi} is the risk-free rate. This can also be written as:

$$R_{p^*} = \alpha + \beta R_{m^*} + \epsilon_i$$

so that R_{p^*} is the excess return on the portfolio and R_{m^*} is the excess return on the market.

Note that both coefficients are important to practitioners. The slope β measures the sensitivity to changes in the market and the contribution of the market premium $E[R_{m^*}]$ to the overall return on the portfolio because:

$$E[R_{p^*}] = \alpha + \beta E[R_{m^*}] \qquad (7.20)$$

Table 7.1 The First Column of Numbers Reports the Estimated Value of the Regression Parameter. The Second Column Reports the Estimated Standard Error. The Third Column Reports the t-Statistic, Which Is the Ratio of the First Two Columns. The Column Labeled p-value Contains the p-value of the t-Statistic and the Column Labeled Confidence Interval Contains a 95% Confidence Interval for the Parameter. The Final Column Reports the R^2 for the Model

	Parameter	Estimate	Std. Err.	t-stat.	p-value	Conf. Int.	R^2
Banking	α	0.026	0.208	0.124	0.901	[−0.382, 0.433]	0.584
	β	1.100	0.049	22.49	0.000	[1.004, 1.196]	
Beer and Liquor	α	0.515	0.220	2.339	0.019	[0.083, 0.946]	0.251
	β	0.569	0.052	10.98	0.000	[0.467, 0.670]	
Chemicals	α	0.054	0.183	0.298	0.766	[−0.304, 0.412]	0.618
	β	1.038	0.043	24.16	0.000	[0.953, 1.122]	
Computers	α	−0.104	0.269	−0.386	0.700	[−0.631, 0.424]	0.581
	β	1.414	0.063	22.34	0.000	[1.290, 1.538]	
Consumer Goods	α	0.245	0.167	1.461	0.144	[−0.084, 0.573]	0.418
	β	0.632	0.039	16.06	0.000	[0.555, 0.709]	
Electrical Equipment	α	0.130	0.177	0.738	0.460	[−0.216, 0.477]	0.713
	β	1.244	0.042	29.93	0.000	[1.162, 1.325]	
Retail	α	0.224	0.164	1.368	0.171	[−0.097, 0.545]	0.613
	β	0.920	0.039	23.89	0.000	[0.845, 0.996]	
Shipping Containers	α	0.062	0.227	0.273	0.785	[−0.384, 0.508]	0.480
	β	0.975	0.053	18.23	0.000	[0.870, 1.080]	
Transportation	α	0.091	0.176	0.517	0.605	[−0.254, 0.436]	0.593
	β	0.948	0.041	22.89	0.000	[0.867, 1.029]	
Wholesale	α	−0.043	0.140	−0.308	0.758	[−0.318, 0.232]	0.667
	β	0.885	0.033	26.86	0.000	[0.821, 0.950]	

The intercept α is a measure of the abnormal return. This coefficient measures the return that the portfolio generates in excess of what would be expected given its exposure to the market risk premium $\beta E[R_{m^*}]$.

As an illustration, CAPM is estimated using 30 years of monthly data between 1989 and 2018. The portfolios measure the value-weighted return to all firms in a sector. The sectors include banks, technology companies, and industrials. The market portfolio measures the return on the complete US equity market. The risk-free rate proxy is the interest rate of a one-month US T-bill.

Table 7.1 reports the parameter estimates, standard errors, t-statistics, p-values and 95% confidence intervals for the estimated parameters. The CAPM βs range from 0.569 for the beer and liquor sector to 1.414 for the computers sector. All estimates of β are statistically different from 0 since their t-ratios are all >1.96 and therefore the p-values are all <0.05.

The hypothesis $H_0{:}\alpha = 0$ can be equivalently tested using any of the three reported measures: the t-statistic, the p-value, or the confidence interval. The t-statistic is the ratio of the estimated parameter to its standard error. The standard error of $\hat{\alpha}$ is estimated using

$$\text{s.e.}(\hat{\alpha}) = \sqrt{\frac{s^2(\hat{\mu}_X^2 + \hat{\sigma}_X^2)}{n\hat{\sigma}_X^2}} = \sqrt{\frac{s^2\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x_i - \bar{X})^2}}, \quad (7.21)$$

where $\hat{\mu}_X$ is the mean of X and $\hat{\sigma}_X^2 = n^{-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$.

When using the t-statistic, values larger than 1.96 indicate that the null should be rejected. If using the p-value, values smaller than 0.05 indicate rejection of the null. Finally, when using the confidence interval, the null is rejected when 0 is not in

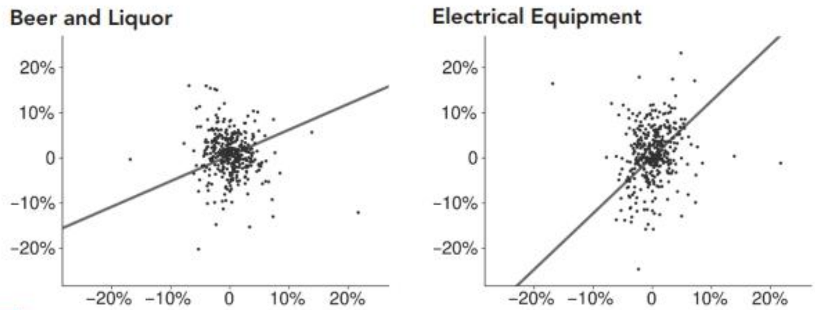


Figure 7.5 Plot of the monthly portfolio returns on the beer and liquor and electrical equipment sectors against the market (x-axis) between 1989 and 2018. The line is the fitted regression.

the confidence interval. The estimates of α here are generally close to 0 and are not statistically different from 0 in 8 of the 9 portfolios.

The final column of Table 7.1 reports the R^2 of the model. This is a measure of the fit of the model and reports the percentage of the total variation in R_{p^*} that can be explained by R_{m^*}. In linear regressions with a single explanatory variable, the R^2 is equal to the squared sample correlation between the dependent and the explanatory variable:

$$R^2 = \text{Corr}^2[R_p^e, R_m^e] \quad (7.22)$$

It also can be used to understand how the variance of the model error is related to the variance of the dependent variable, because:

$$\sigma^2 = (1 - R^2)\sigma_Y^2$$

This measure is more useful when examining models with more than one explanatory variable, as shown in Chapter 8. In this application, it is useful to determine which sectors are more closely coupled to the market. For example, electrical equipment and wholesale both have large R^2, which indicates that the returns of these sectors are tightly coupled with the market return (i.e., a higher proportion of the total variability of the returns of those sectors can be attributed to general equity market movements).

Figure 7.5 plots the excess market returns and the excess returns for electrical equipment and beer and liquor. Overall, the patterns among the returns do not look very different. However, the beer and liquor portfolio had a large negative return when the market had its largest positive return in this sample, which lowers the slope and reduces the R^2.

7.6 APPLICATION: HEDGING

Estimating the optimal hedge ratio is a natural application of linear regression. Hedges are commonly used to eliminate undesirable risks. For example, a long-short equity hedge fund attempts to profit from differences in returns between firms, but not on the return to the overall market. However, the composition of the fund's long and short positions may produce an unwanted correlation between the fund's return and the market if the market exposures of the fund's positions do not fully offset each other. A hedge can be used to remove the market risk.

In this case, the slope coefficient is directly interpretable as the hedge ratio that minimizes the variance of the unhedged risk.[8] The regression model estimated is:

$$R_{pi} = \alpha + \beta R_{hi} + \epsilon_i, \quad (7.23)$$

where R_{pi} is the return on the portfolio to be hedged and R_{hi} is the return on the hedging instrument. The β is the optimal hedge ratio and α measures the expected return on the hedged portfolio (i.e., the return on the original portfolio minus the return of the hedging instrument scaled by the hedge ratio).

Table 7.2 contains estimated hedge ratios for a set for widely traded exchange-traded funds (ETFs). These funds are all part of the Select Sector SPDR ETF family and target the return of a diversified industry portfolio. Meanwhile, the hedging instrument is the SPDR S&P 500 ETF, which tracks the return on the S&P 500.

[8] The regression coefficient fully hedges the risk in the sense that the resulting portfolio is uncorrelated with the hedged factor. When factors have positive risk premia, this can be expensive, and it is also common to partially hedge to eliminate some, but not all, factor risk.

Table 7.2 Optimal Hedge Ratios for Select Sector SPDR ETFS Hedged against the Market Using the SPDR S&P 500 ETF. The First Column, μ, Reports the Annualized Mean Return of the ETF. The Column Labeled α Reports the Estimate of the Annualized Return on the Hedged Portfolio. The Third Column Is the Estimate of the Hedge Ratio. The Columns s_p and s_h Report the Standard Deviation of the ETF Return and of the Hedged Return, Respectively. The Final Column Reports the R^2 from the Model Used to Estimate the Hedge Ratio. The t-Statistics Are Reported in Parentheses Below the Estimated Parameters

	μ	α	β	s_p	s_h	R^2
Consumer	6.235	2.907	0.483	12.0	9.8	0.335
	(2.30)	(1.30)	(10.81)			
Energy	10.188	3.907	0.913	21.3	16.8	0.377
	(2.11)	(1.01)	(11.85)			
Financial	7.480	−0.770	1.199	22.0	13.7	0.613
	(1.50)	(−0.25)	(19.15)			
Health	8.298	3.129	0.751	13.7	8.5	0.617
	(2.67)	(1.61)	(19.33)			
Materials	9.442	1.403	1.168	21.0	12.7	0.634
	(1.98)	(0.48)	(20.07)			
Technology	7.713	−1.682	1.365	23.3	12.6	0.706
	(1.46)	(−0.58)	(23.63)			
Utilities	7.473	4.481	0.435	14.9	13.5	0.176
	(2.22)	(1.45)	(7.04)			

The models are estimated using monthly data from 1999 until 2018. The first column reports the average return on the unhedged fund. The next two report the estimates of the regression coefficients. The t-statistics are reported below each coefficient. The estimate of α has been multiplied by 12 so that it can be interpreted as the annual hedged portfolio return. Scaling α has no effect on the t-statistic, and so hypothesis tests are similarly unaffected by this type of rescaling.

Note that all unhedged returns are positive, and five are statistically different from zero. The hedged returns, however, have mixed signs and are not statistically different from zero when using a 5% test (i.e., their t-statistics are less than 1.96 in absolute value). The third column is the estimate of the hedge ratio. The next two columns report the standard deviations of the unhedged (s_p) and hedged portfolios (s_h). The quantity s_h is the same as the standard error of the regression (s), which measures the standard deviation of the residuals in the model. It can be seen that the standard deviation of a hedged portfolio is often much lower than that of its unhedged counterpart. Note, however, that this has come at the expense of much lower returns (the μ are all positive and statistically significant while the α, the hedged portfolio returns, are not). Furthermore, this reduction in standard deviation is largest for the models with the highest R^2.

7.7 APPLICATION: PERFORMANCE EVALUATION

Fund managers are frequently evaluated against a style benchmark to control for the typical performance of a fund with the same investment style. High-performing fund managers should outperform their benchmarks and thus generate positive α. Performance analysis can be implemented using a regression where the return on the fund is regressed on the return to its benchmark. α is the most important parameter in this application because it can be used to detect superior or inferior performance; β measures the sensitivity to the benchmark and so can be used to examine whether the benchmark is appropriate for the fund's return. A good choice of a benchmark should have a β near one and a large R^2 because $\beta \equiv 1$ indicates that the risks captured by the benchmark are similar to the fund's and large R^2 indicates that the benchmark is highly correlated with the fund's performance.

Table 7.3 The Left Panel Reports $\hat{\alpha}$, $\hat{\beta}$, and the R^2 from a Performance Evaluation Model Using the Style Benchmark Listed in Italics. The Right Column Uses a Market Model where the Benchmark Is the Return on the S&P 500. The t-Statistics Are Reported Below Estimated Parameters. The Final Column Reports the Number of Observations Available to Estimate Parameters

	Style			Market			
	α	β	R^2	α	β	R^2	n
Fidelity Contrafund	4.072	0.861	0.836	5.639	0.955	0.803	619
Large Growth	(4.47)	(56.04)		(5.68)	(50.16)		
JPMorgan Small Cap Growth A	1.075	0.933	0.912	3.012	1.094	0.602	329
Small Growth	(0.94)	(58.05)		(1.24)	(22.25)		
PIMCO Total Return Fund	1.172	1.055	0.891	6.565	0.050	0.031	379
Intermediate-Term Bond	(4.42)	(55.41)		(8.94)	(3.47)		
PIMCO High Yield Fund	1.713	0.919	0.902	4.864	0.326	0.409	312
High Yield Bond	(3.73)	(53.29)		(4.39)	(14.65)		

The choice of benchmark plays an important role when assessing whether a fund produces excess returns. If a fund invests in risks that are not captured by the benchmark, then the fund might appear to produce α (because the benchmark cannot account for all risk exposures). In practice, it is common to use multiple factors—between 3 and 11—depending on the fund type—to fully account for portfolio risk exposures. This is the topic of the next chapter.

Table 7.3 contains results from the regression:

$$R_{FUND} = \alpha + \beta R_{si} + \epsilon, \quad (7.24)$$

where R_{FUND} is the monthly return on the fund and R_s is the return on the style benchmark. The performance of four funds, each using a different investment style, is examined. Two of the funds, the Fidelity Contrafund and the JPMorgan Small Cap Growth Fund, invest in high-growth equities. The Fidelity fund invests in large-cap stocks, while the JPMorgan fund focuses on smaller firms. The other two funds, the PIMCO Total Return Fund and the PIMCO High Yield Fund, invest in bonds. The total return fund targets the total return to investing in US bonds, while the high-yield fund focuses on generating returns from non-investment grade securities. Each of the four funds has a different style, and so the benchmark is adjusted to match a given fund's investment objective.

These regressions use monthly returns, and the reported excess performance is annualized by multiplying by 12. All monthly returns available for each fund, from the fund's launch until the end of 2018, are used in the analysis. The number of returns used to estimate the parameters in each model is reported in the table.

Estimates of α are greater than zero for all four funds, and three of these are statistically different from zero. Fidelity's Contrafund has outperformed its benchmark by over 4% per year during the 50 years in the sample.[9] However, the Contrafund also has the smallest β estimate and R^2. This suggests that the benchmark portfolio might not be appropriate for the strategy used by this fund. For the other three funds, the R^2 values show that the benchmarks are highly correlated with the fund returns. The $\hat{\beta}$ are all close to 1, which also indicates that the funds' returns closely comove with their benchmarks.

The right column repeats the exercise using the return on the S&P 500 benchmark. The market appears to be nearly as good as the large growth style return in adjusting the return on Fidelity's Contrafund, and the two R^2 values are similar. The estimated α is somewhat larger and is still significantly larger than 0. It has a lower correlation with the small-cap fund because the market portfolio is, by construction, a large-cap portfolio. The conclusion is, however, the same, and the α is not statistically different from 0.

The R^2 on the bond funds are both lower, and near 0 for the PIMCO Total Return Fund. The poor fit of the market model indicates that the return on the equity market is not a good benchmark to use when assessing the performance of these bond funds.

[9] These funds have traded for more than 25 years. The estimates of α are not representative because this sample suffers from survivorship bias. The funds have all survived for an extended period which is only likely if a fund meets or exceeds its performance target.

7.8 SUMMARY

This chapter introduces linear regression through a focus on models that contain a single explanatory variable. Parameter estimators in linear regression models have simple closed-forms that depend on only the first two moments of Y and X. The slope parameter is closely related to the correlation between the dependent and explanatory variables, and the slope is zero if and only if the two variables are uncorrelated.

Deriving the estimators for these only requires one trivial assumption—that the explanatory variable has some variation. Interpreting the results of an estimated regression requires four additional assumptions. The most important of these, and the most difficult to verify, is the conditional mean assumption:

$$E[\epsilon_i | X_i] = 0$$

This assumption rules out some empirically relevant applications (e.g., when the sample has selection bias or if X and Y are simultaneously determined). The other assumptions (i.e., that the data are generated by iid random variables, have no outliers, and are homoscedastic) are simpler to verify using either graphical analysis or formal tests.

When these assumptions are satisfied, then β can be interpreted as the change in Y per unit change in X. The OLS estimators of α and β also satisfy a CLT and so have a normal distribution in large samples. The CLT allows hypothesis testing using t-tests and for confidence intervals to be constructed. The t-statistic, a t-test for the null hypothesis that the parameter is zero, is frequently reported to indicate whether a parameter is statistically significant. Finally, the chapter presents three standard applications of regression: factor sensitivity estimation, hedging, and fund performance evaluation.

The following questions are intended to help candidates understand the material. They are *not* actual FRM exam questions.

QUESTIONS

Short Concept Questions

7.1 What are the three requirements of linear regression?

7.2 You suspect that the CAPM held on all days except those with a Federal Open Markets Committee (FOMC) announcement, and on these days the β is different. How can a dummy be used to capture this effect? What could you do if you suspected that both α and β are different on FOMC days?

7.3 What does a confidence interval for a regression parameter measure?

7.4 How is a *t*-test related to the *t*-statistic of a parameter?

7.5 The R^2 from regressing Y on X is the same as the R^2 of regressing X on Y. True or false?

7.6 What is R^2 if $\hat{\beta} = 0$?

7.7 The return on an optimally hedged portfolio is independent of the return of the hedging instrument. True or false?

7.8 What are the consequences of using a poor benchmark to evaluate a fund manager?

Practice Questions

7.9 Find the OLS estimates for the following data:

x	y
0	−1.46
1	0.35
2	6.46
3	4.09
4	7.34
5	6.18
6	14.97
7	14.28
8	20.20
9	21.24

7.10 In running a regression of the returns of Stock XYZ against the returns on the market, the standard deviation of the returns of Stock XYZ is 20%, and that of the market returns is 15%. If the estimated beta is found to be 0.75:

a. What is the correlation between the returns of Stock XYZ and those of the market?

b. If the market falls by 2%, what is the expected return on Stock XYZ?

c. What is the maximum possible value of beta given that the standard deviation of the returns of Stock XYZ is 20% and those of the market is 15%?

7.11 In a CAPM that regresses Wells Fargo's excess returns on the market's excess returns, the coefficients on monthly data are $\alpha = 0.1$ and $\beta = 1.2$. What is the expected excess return on Wells Fargo when the excess return on the market is 3.5%?

7.12 You fit a CAPM that regresses the excess return of Coca-Cola on the excess market return using 20 years of monthly data. You estimate $\hat{\alpha} = 0.71$, $\hat{\beta} = 1.37$, $s^2 = 20.38$, $\hat{\sigma}^2_X = 19.82$ and $\hat{\mu}_X = 0.71$.

a. What are the standard errors of $\hat{\alpha}$ and $\hat{\beta}$?

b. What are the *t*-statistics for $\hat{\alpha}$ and $\hat{\beta}$?

c. What is the 99% confidence interval for $\hat{\beta}$?

7.13 Suppose that you estimated a model that regressed the volume of put and call options traded on the S&P 500 on a measure of overnight news, $VOL_i = \alpha + \beta NEWS_i + \epsilon_i$. Your statistical package returned a 90% confidence interval for β of [0.32,1.89].

a. Is β statistically significant when using a 5% test?

b. What is the p-value of the *t*-statistic of β?

7.14 If $\hat{\beta} = 1.05$ and its 95% confidence interval is [0.43,1.67], what is the *t*-statistic of $\hat{\beta}$? What is the p-value of the *t*-statistic?

7.15 How does adding leverage (borrowing at a risk-free rate, which is not random) to a portfolio change the optimal hedge ratio? For example, if the portfolio returns were doubled so that $\tilde{R}_p = 2R_p$, what is the optimal hedge ratio for the R_p? Hint: Use the formula for the OLS estimator of β.

The following questions are intended to help candidates understand the material. They are ***not*** actual FRM exam questions.

ANSWERS

Short Concept Questions

7.1 The three requirements are that: the model is linear in the unknown coefficients, the error is additive, and there are no missing values.

7.2 The model that allows differences in the slope would be $R_i = \alpha + \beta R_{m,i} + \gamma I_{FOMC} R_{m,i} + \epsilon_i$ where I_{FOMC} is a dummy variable taking the value 1 on FOMC days and 0 otherwise. If γ is not zero, then the slope is different on FOMC days. This can be extended to both parameters by estimating the model with both an intercept and a slope dummy variable:

$$R_i = \alpha + \gamma_1 I_{FOMC} + \beta R_{m,i} + \gamma_2 I_{FOMC} R_{m,i} + \epsilon_i$$

7.3 The $1 - \alpha$ confidence interval contains the set of null hypotheses that could not be rejected using a test size of α.

7.4 The t-stat is a t-test of the null $H_0: \beta = 0$ against the alternative $H_1: \beta \neq 0$.

7.5 True. The R^2 in a regression with a single explanatory variable is the squared sample correlation between X and Y.

7.6 The only time $\hat{\beta} = 0$ is when $\widehat{Cov}(X,Y) = 0$ and so the two variables are uncorrelated. The R^2 is the squared correlation and so must be 0.

7.7 False. The optimally hedged portfolio is uncorrelated with the return on the hedge (i.e., not linearly related). It is not necessarily independent. It would be independent if the returns were jointly normally distributed.

7.8 If the benchmark is poor, so that the evaluation model is misspecified, then the slope β will be mismeasured. As a result, some of the compensation for systemic risk (which is captured by β) may be attributed to skill α. In other words, we could be misled into believing that the manager is out-performing when in reality he or she is just receiving fair compensation for taking other risks not incorporated in the model.

Solved Problems

7.9 Doing the basic calculations needed:

	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(x-\bar{x})(y-\bar{y})$	$(y-(\hat{\alpha}+\hat{\beta}x))^2$
	0	−1.46	−4.50	−10.86	20.25	48.85	0.25
	1	0.35	−3.50	−9.04	12.25	31.66	0.04
	2	6.46	−2.50	−2.93	6.25	7.33	11.41
	3	4.09	−1.50	−5.30	2.25	7.96	2.31
	4	7.34	−0.50	−2.05	0.25	1.03	0.62
	5	6.18	0.50	−3.22	0.25	−1.61	20.06
	6	14.97	1.50	5.57	2.25	8.36	3.19
	7	14.28	2.50	4.89	6.25	12.22	2.03
	8	20.20	3.50	10.80	12.25	37.82	3.88
	9	21.54	4.50	12.14	20.25	54.64	0.61
Average	*4.50*	*9.40*			*8.25*	*20.83*	
SUM					*82.50*	*208.25*	*44.41*

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{20.83}{8.25} = 2.52$$

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X} = 9.40 - 2.52 * 4.50 = -1.96$$

Continuing to estimate the variance of the errors:

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\epsilon}_i^2 = \frac{1}{10-2}19.87 = 5.55$$

7.10 **a.** $\hat{\beta} = \frac{\hat{\rho}_{xy}\hat{\sigma}_y}{\hat{\sigma}_x} \rightarrow 0.75 = \frac{\hat{\rho}_{xy}0.2}{0.15} \rightarrow \hat{\rho}_{xy} = 56.25\%$

b. $\hat{\beta} * 2\% = 0.75 * 2\% = 1.5\%$

c. The maximum correlation is 1, therefore:

$$\hat{\beta}_{max} = \frac{1 * \hat{\sigma}_y}{\hat{\sigma}_x} = \frac{0.20}{0.15} = 1.33$$

7.11 The expected return is $0.1 + 1.2 \times 3.5\% = 14.2\%$. This value is the fitted value from the linear regression when the market return is 3.5%.

7.12 **a.** The standard error for $\hat{\beta}$ is

$$\sqrt{\frac{s^2}{n \times \hat{\sigma}_X^2}} = \sqrt{\frac{20.38}{240 \times 19.82}} = 0.065.$$

The standard error for $\hat{\alpha}$ is

$$\sqrt{\frac{s^2 \times (\hat{\mu}_X^2 + \hat{\sigma}_X^2)}{n\hat{\sigma}_X^2}} = \sqrt{\frac{20.38 \times (0.71^2 + 19.82)}{240 \times 19.82}} = 0.295.$$

b. The t-stat for $\hat{\beta}$ is $\frac{1.37}{0.065} = 20.9$.

The t-stat for $\hat{\alpha}$ is $\frac{0.71}{0.295} = 2.40$.

c. A 99% two-sided confidence interval uses the critical value for a test with a size of 1% (0.5% in each tail), which is 2.57. The confidence interval is then $[1.37 - 2.58 \times 0.065, 1.37 + 2.58 \times 0.065] = [1.20, 1.54]$.

7.13 **a.** Yes. The null hypothesis value of $H_0: \beta = 0$ is not in the confidence interval, and so the null is rejected using a test size of 10%.

b. The range of the confidence interval is $2c \times se(\hat{\beta})$, where c is the critical value used to construct the 90% confidence interval (5% in each tail, so 1.645). The estimate of $\hat{\beta}$ is the midpoint of the confidence interval so 1.105. Because $\hat{\mu} - 1.645se(\hat{\beta}) = 0.32$ and $\hat{\mu} + 1.645se(\hat{\beta}) = 1.89$, these can be solved for $se(\hat{\beta}) = 0.477$. The t-stat is then $\frac{\hat{\beta}}{se(\hat{\beta})} = 2.32$. Finally, the two-sided p-value is $2(1 - \Phi(|t|)) = 2.06\%$, where $\Phi(\cdot)$ is the standard normal CDF function. This p-value confirms the rejection in the previous step because it is less than 10%.

7.14 The standard error is $\frac{1.67 - 0.43}{2 \times 1.96} = 0.316$.

The t-stat is then 3.31. The two-sided p-value is $2(1 - \Phi(|t|)) = 0.09\%$.

7.15 The OLS β estimator is $\frac{\widehat{Cov}(R_p, R_h)}{\widehat{Var}(R_h)}$. Scaling R_p by a leverage v, the optimal hedge would be $\frac{\widehat{Cov}(vR_p, R_h)}{\widehat{Var}(R_h)} = v\frac{\widehat{Cov}(R_p, R_h)}{\widehat{Var}(R_h)} = v\hat{\beta}$, where $\hat{\beta}$ is the ledge ratio from the unlevered portfolio. Levering up increases exposure to risk that can be hedged and so the hedge ratio must account for this.

8

Regression with Multiple Explanatory Variables

Learning Objectives

After completing this reading you should be able to:

- Distinguish between the relative assumptions of single and multiple regression.
- Interpret regression coefficients in a multiple regression.
- Interpret goodness-of-fit measures for single and multiple regressions, including R^2 and adjusted R^2.
- Construct, apply, and interpret joint hypothesis tests and confidence intervals for multiple coefficients in a regression.

Linear regression with a single explanatory variable provides key insights into OLS estimators and their properties. In practice, however, models typically use multiple variables where it is possible to isolate the unique contribution of each explanatory variable. A model built with multiple variables can also distinguish the effect of a novel predictor from the set of explanatory variables known to be related to the dependent variable.

This chapter introduces the *k*-variate regression model, which enables the coefficients to measure the distinct contribution of each explanatory variable to the variation in the dependent variable. This structure allows the estimated parameters to be interpreted while holding the value of other included variables constant.

The first section of this chapter shows how the structure of the OLS estimator leads to this interpretation in a model with two correlated explanatory variables. The coefficients in this model can be estimated using three single-variable regressions. This multistep estimation procedure provides an intuitive understanding of the effect measured by a regression coefficient.

R^2, introduced in the previous chapter, is a simple measure that describes the amount of variation explained by a model. It is also the basis of the *F*-test, which extends the *t*-test to allow multiple hypotheses (or hypotheses involving more than one parameter) to be tested simultaneously. The *F*-test accounts for the uncertainty in the coefficient estimators, including the correlation between them.

The key ideas in this chapter are illustrated through an extension of CAPM, which incorporates additional risk factors that systematically affect the cross section of asset returns. These multi-factor models are commonly used in risk management to determine sensitivities to economically important investment characteristics and to risk-adjust returns when assessing fund manager performance.

8.1 MULTIPLE EXPLANATORY VARIABLES

Linear regression can be directly extended to models with more than one explanatory variable.[1] The general structure of such a regression is:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i, \quad (8.1)$$

where there are *k* explanatory variables and a single constant α. Each explanatory variable now has two subscripts: the variable index and the observation index. For example, X_{2i} is the i^{th} observation on the second variable.

[1] Historically, models with multiple explanatory variables are called multiple regressions. This distinction is not important and so models with any number of explanatory variables are referred to as linear regressions or simply regressions.

REGRESSION WITH INDISTINCT VARIABLES

If some explanatory variables are functions of the same random variable (e.g., if $X_2 = X_1^2$), then:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \varepsilon_i$$

In this case, it is not possible to change X_1 while holding the other variables constant. The interpretation of the coefficients in models with this structure depends on the value of X_1 because a small change of ΔX_1 in X_1 changes Y by $(\beta_1 + 2\beta_2 X_1)\Delta X_1$. This effect captures the direct, linear effect of a change in X_1 through β_1 and its nonlinear effect through β_2.

When all explanatory variables are distinct (i.e., no variable is an exact function of the others), then the coefficients are interpreted as *holding all other values fixed*. For example, β_1 is the effect of a small increase in X_1 holding all other variables constant.

While the parameter estimators of α and the βs have closed forms, they can be difficult to express without matrices and linear algebra. However, it is possible to understand how parameters are estimated in a model with two explanatory variables. Suppose the model is

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

The OLS estimator for β_1 can be computed using *three* single-variable regressions. The first regresses X_1 on X_2 and retains the residuals from this regression:

$$\tilde{X}_{1i} = X_{1i} - \hat{\delta}_0 - \hat{\delta}_1 X_{2i}, \quad (8.2)$$

where $\hat{\delta}_0$ and $\hat{\delta}_1$ are OLS parameter estimators and $\tilde{X}_1$ is the residual of X_1.

The second regression does the same for Y:

$$\tilde{Y}_i = Y_i - \hat{\gamma}_0 - \hat{\gamma}_1 X_{2i}, \quad (8.3)$$

where $\hat{\gamma}_0$ and $\hat{\gamma}_1$ are OLS parameter estimators, and $\tilde{Y}_1$ is the residual of Y_1

The final step regresses the residual of Y (i.e., $\tilde{Y}$) on the residual of X_1 (i.e., $\tilde{X}_1$):

$$\tilde{Y}_i = \hat{\beta}_1 \tilde{X}_{1i} + \varepsilon_i \quad (8.4)$$

The OLS estimate $\hat{\beta}_1$ is identical to the estimate computed by fitting the full model using Excel or any other statistical software package.[2]

The first two regressions have a single purpose: to remove the direct effect of X_2 from Y and X_1. They do this by decomposing

[2] This model does not include a constant because $\tilde{Y}$ and $\tilde{X}$ have mean zero by construction. If a constant is included, its estimated value is exactly 0 and the estimate of β_1 is unaffected.

each variable into two components: one that is perfectly correlated with X_2 (i.e., the fitted value) and one that is uncorrelated with X_2 (i.e., the residual). As a result, the two residuals (i.e., $\tilde{Y}$ and $\tilde{X}_1$) are uncorrelated with X_2 by construction.

The final regression estimates the linear relationship (i.e., β_1) between the components of Y and X_1 that are uncorrelated with (and so cannot be explained by) X_2.

Finally, the OLS estimate of β_2 can be computed in the same manner by reversing the roles of X_1 and X_2 (i.e., so that β_2 measures the effect of the component in X_2 that is uncorrelated with X_1).

This multistep estimation can be used to estimate models with any number of regressors. In the k-variable model:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \varepsilon_i, \qquad (8.5)$$

the OLS estimate of β_1 could be computed by first regressing each of X_1 and Y on a constant and the remaining $k - 1$ explanatory variables. The residuals from these two regressions are mean zero and uncorrelated with the remaining $k - 1$ explanatory variables. The OLS estimator of β_1 is then estimated by regressing the residuals of Y on the residuals of X_1.

Additional Assumptions

Extending the model to multiple regressors requires one additional assumption, along with some modifications, to the five assumptions in Section 7.3.

Multiple linear regression assumes that the explanatory variables are not perfectly linearly dependent (i.e., each explanatory variable must have some variation that cannot be perfectly explained by the other variables in the model).

The distinct variation in each regressor is exploited to estimate the model parameters. If this assumption is violated, then the variables are *perfectly collinear*. This assumption is simple to verify in practice, and most statistical software packages produce a warning or an error when the explanatory variables are perfectly collinear.

The remaining assumptions require simple modifications to account for the k-explanatory variables.

- All variables must have positive variances so that $\sigma^2_{X_j} > 0$ for $j = 1, 2, \dots, k$.
- The error is assumed to have mean zero conditional on the explanatory variables ($E[\varepsilon_i | X_{1i}, X_{2i}, \dots, X_{ki}] = 0$).
- The random variables ($X_{1i}, X_{2i}, \dots, X_{ki}$) are assumed to be iid.
- The assumption of no outliers is extended to hold for each explanatory variable so that $E[X^4_{ji}] < \infty$ for $j = 1, 2 \dots, k$.
- The constant variance assumption is similarly extended to hold for all explanatory variables ($E[\varepsilon_i^2 | X_{1i}, X_{2i}, \dots, X_{ki}] = \sigma^2$).

The interpretation of each of the modified assumptions is unchanged from the single explanatory variable model.

REGRESSION STEP-BY-STEP

The Appendix contains a table of the annual excess returns on the shipping container industry portfolio (i.e., R_{p^*}), the excess return on the market (i.e., R_{m^*}), and the return on a value portfolio (i.e., R_v).

It also contains the squared deviations from the mean of each variable (labeled E_p^2, E_m^2, and E_v^2, for squared error) and the cross-products of the deviations (E_pE_m and E_vE_m). The sums of these columns, when divided by the sample size, are the estimates of the variances (squares) and covariances (cross-products) of these series.

If this data set were to be entered into Excel and a regression run using the Analysis ToolPak add-in, the fitted model would be

$$\hat{R}_{p^*} = \underset{(4.16)}{4.06} + \underset{(0.22)}{0.72} R_{m^*} + \underset{(0.29)}{0.101} R_v + \hat{\varepsilon}, \quad R^2 = 0.394,$$

where the standard error is in parentheses below each coefficient.

Here the multistep procedure is applied to verify the estimate of the coefficient on the value factor.

The first step is to estimate the coefficients of the model that removes the effect of the market from the shipping container portfolio:

$$R_{p^*} = \delta_0 + \delta_1 R_{m^*} + \eta_p$$

The slope (i.e., $\hat{\delta}_1$) depends on the covariance between the industry portfolio and the market portfolio divided by the variance of the market portfolio. This is estimated by using the sum of the cross-product between the two and the sum of the squared market return deviations. (Note that it is unnecessary to divide each by $n - 1$ to produce a covariance or variance estimate, because this cancels in the ratio.). The intercept depends on the estimate $\hat{\delta}_1$ and the means of the two series:

$$\hat{\delta}_1 = \frac{4639}{6665} = 0.696$$

The intercept depends on the estimate $\hat{\delta}_1$ and the means of the two series:

$$\hat{\delta}_0 = 8.58 - 0.696 \times 6.03 = 4.39$$

(Continued)

(Continued)

These same calculations are used to estimate the regression of the value factor return on the excess return of the market:

$$R_v = \gamma_0 + \gamma_1 R_{m^*} + \eta_v$$

The parameter estimates are:

$$\hat{\gamma}_1 = -\frac{1464}{6665} = -0.220,\ \hat{\gamma}_0 = 1.94 - (-0.220) \times 6.03 = 3.27$$

Finally, the effect of the market is removed using the estimated coefficients to produce the following residuals:

$$\tilde{R}_{p^*i} = R_{p^*i} - 4.39 - 0.696\, R_{m^*i}$$
$$\tilde{R}_{vi} = R_{vi} - 3.27 - (-0.220)\, R_{m^*i}$$

The values of this residual series are in the second table in the appendix. The first two columns sum to exactly 0, which is a consequence of fitting the first step model. The final three columns contain the squares and cross-products of the residual series. These are the key inputs into the coefficient estimate for the value factor, which depends on the covariance between the two residuals divided by the variance of the value factor residual. As before, the $n - 1$ term can be omitted in each because it cancels, and the estimate of the regression coefficient is

$$\hat{\beta}_v = \frac{353}{3483} = 0.101$$

8.2 APPLICATION: MULTI-FACTOR RISK MODELS

Market variation is an important determinant of the variation in asset returns. However, other factors that capture specific characteristics are also useful in explaining differences in returns across firms or industries.

The Fama-French three-factor model[3] is a leading example of a multi-factor approach. This model expands upon CAPM by including two additional factors: the size factor (which captures the propensity of small-cap stocks to generate higher returns than large-cap stocks) and the value factor (which measures the additional return that value[4] stocks earn above growth stocks).

Both the size and value factors are measured using the returns of long-short portfolios that buy stocks with the desirable feature (i.e., small-cap or value stocks) and short sell those with the undesirable feature (i.e., large-cap or growth[5] stocks).

Table 8.1 extends the results for CAPM in Table 7.1 to the Fama-French three-factor model. The table contains parameter estimates for the model:

$$R_{p^*i} = \alpha + \beta_m R_{m^*i} + \beta_s R_{si} + \beta_v R_{vi} + \varepsilon_i,$$

where R_{m^*} is the excess return on the market, R_{vi} measures the excess return of value firms over growth firms, and R_{si} measures the excess return of small-cap firms over large-cap firms.[6] The returns of all three factors and the industry portfolios are available in Ken French's data library.[7]

The *t*-statistic is reported below each coefficient in parentheses. The coefficients on the market return are all similar in magnitude to those in Table 7.1. They are all positive, close to one, and have large *t*-statistics indicating that the returns on most industries comove closely with the overall market return.

The coefficients on the size and value factors, however, have a different structure. These coefficients are all smaller in magnitude and range from −0.6 to 0.8. Positive estimates indicate that the industry has a positive exposure to the factor. For example, the wholesale industry has a coefficient of 0.26 on the size factor, indicating that a 1% factor return increases the return on this portfolio by 0.26%. This suggests one of two possibilities: either the firms in this industry are relatively small firms, or their profits depend crucially on the same factors as those affecting the performance of small firms. Note that these two explanations for the positive exposure are not exclusive and both are likely true. The computer industry portfolio, on the other hand, has a large negative coefficient on the value factor. This coefficient suggests that the typical firm in this industry is a growth firm.

Figure 8.1 illustrates the difference between estimates of the factor exposure in a single variable model (which only includes the factor return) and in models that include additional explanatory variables.

[3] Fama, E. F., & French, K. R. (1993). Common Risk Factors in the Returns on Stocks and Bonds. Journal of Financial Economics, 33(1), 3–56.

[4] A firm is classified as a value stock if its fundamentals are strong relative to its price. The value ratio is often computed by comparing the book value of assets to the market value of assets (also called the book-to-market ratio). Firms with high book-to-market ratios represent good value because investing in these firms buys more (tangible) assets than an equal investment in non-value firms.

[5] Firms with low book-to-market ratios are called growth firms because the high price of these firms forecasts high growth in their fundamentals (e.g., revenues, profits, or the book value of assets).

[6] The size and value factor returns are not measured in excess of the risk-free rate because their construction as long-short portfolios eliminates the risk-free rate from the portfolio return (i.e., it cancels out).

[7] Ken French's data library can be accessed at: http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

Table 8.1 Regression Results from Models of Industry Portfolio Returns on Three Factors: The Market, the Size Factor, and the Value Factor. The t-Statistics Are Reported in Parentheses Below Each Coefficient. The Right-Most Column Reports the Model R^2

	α	β_m	β_s	β_v	R^2
Banking	−0.172 (−1.09)	1.228 (32.26)	−0.156 (−3.03)	0.838 (15.15)	0.764
Beer and Liquor	0.517 (2.44)	0.628 (12.28)	−0.394 (−5.67)	−0.033 (−0.44)	0.314
Shipping Containers	−0.003 (−0.01)	1.011 (18.63)	−0.012 (−0.16)	0.276 (3.51)	0.498
Chemicals	−0.053 (−0.31)	1.098 (26.90)	−0.028 (−0.51)	0.458 (7.72)	0.677
Electrical Equipment	0.103 (0.59)	1.263 (29.61)	−0.033 (−0.57)	0.115 (1.85)	0.717
Computers	0.055 (0.23)	1.296 (21.99)	0.224 (2.80)	−0.666 (−7.78)	0.659
Consumer Goods	0.215 (1.31)	0.675 (17.03)	−0.174 (−3.24)	0.117 (2.03)	0.446
Retail	0.221 (1.35)	0.927 (23.33)	−0.036 (−0.66)	0.010 (0.17)	0.614
Transportation	−0.012 (−0.07)	0.995 (25.14)	0.045 (0.84)	0.443 (7.71)	0.651
Wholesale	−0.104 (−0.80)	0.877 (27.83)	0.260 (6.09)	0.277 (6.05)	0.715

CONTROL VARIABLES

Controls are explanatory variables that are known to have a clear relationship with the dependent variable. It is frequently the case that the parameter values on the controls are not directly of interest, but still need to be incorporated to ensure that the model includes all relevant factors affecting the dependent variable.

It is common to control for the effects of explanatory variables that are known to have a strong relationship to the dependent variable when examining the explanatory power of a new variable. This practice ensures that the novel variable is finding a feature of the data that is distinct from the effects captured by a standard set of explanatory variables.

Leaving out these controls is known as omitted variable bias. For example, when researchers develop a new measure that explains the cross section of asset prices, it is common to include the market, size, and value factor returns when assessing whether the new measure improves the model.

The top panels contain return plots of the computer industry portfolio against the return on the size factor (left) or the return on the value factor (right). The solid line shows the fitted regression of the excess industry portfolio return on the factor return. For example, the model fit in the top left panel is:

$$R_{p^*i} = \alpha + \beta_s R_{si} + \varepsilon_i$$

The size factor has a strong positive relationship (0.762) to the computer industry return, and the value factor model indicates a strong negative relationship (−1.04). These estimates differ from those included in Table 7.1 because two variables (i.e., the market and the other factor) are excluded from the model.

The middle panel shows the effect of controlling for the market return. This is done by plotting:

- The residuals calculated by regressing the sector portfolio returns on the market (i.e., the y-axis), and
- The residuals calculated by regressing the factor being examined on the market (i.e., the x-axis).

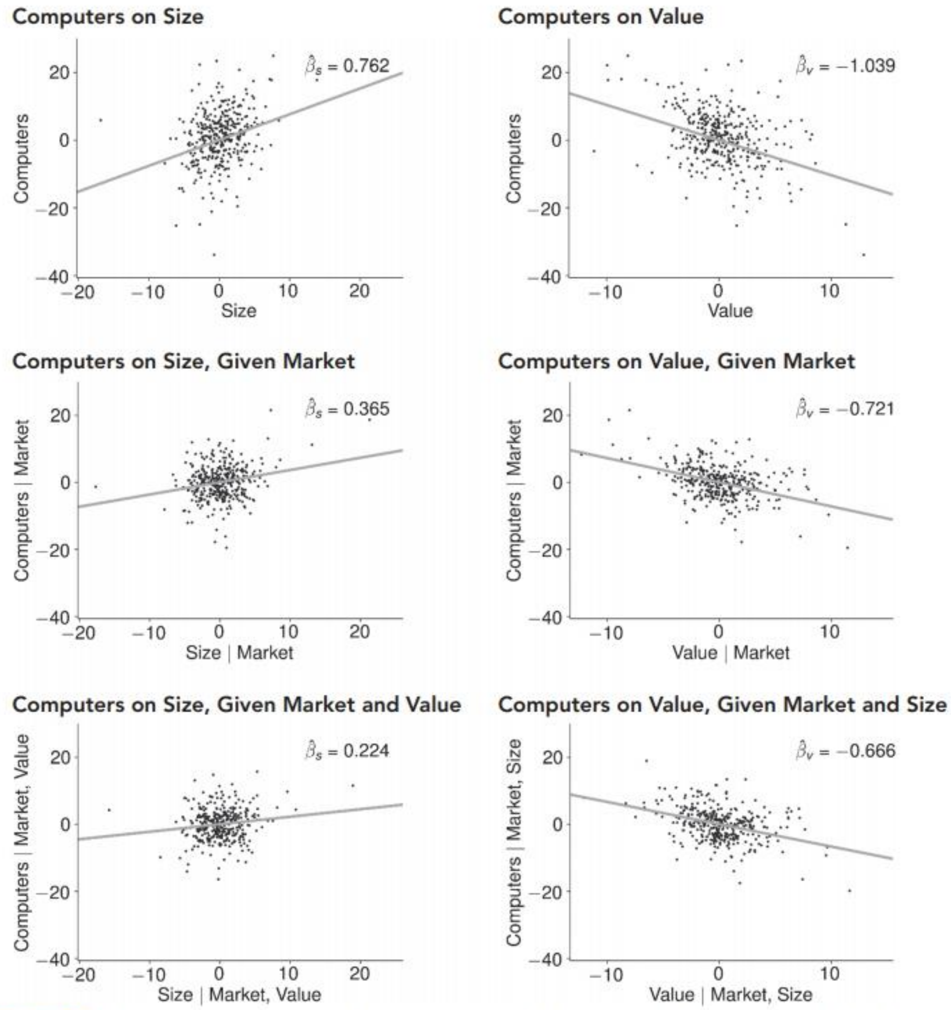


Figure 8.1 Illustration of the differences between estimates of the direct effect of a variable and the effect given other relevant variables. Each column contains plots that depict the relationship between the computer industry returns and either the size factor (left) or the value (right) factor. The top panel in each column plots the portfolio returns (y-axis) against the factor returns and shows the regression line computed by regressing the portfolio return on only the factor return. The middle panel shows the relationship between the portfolio return and the factor return when controlling for the market return. The values plotted are the residuals from regressing the portfolio return on the market return (y-axis) and the factor return on the market return (x-axis). The bottom panel shows the relationship when controlling for both the market and the other factor. The values shown are also residuals where the effect of the market and the other factor are eliminated.

Controlling for the market removes its direct effect from both the industry portfolio return and the factor return. The slope $\hat{\beta}_s$ is then estimated in a regression that includes either the size or the value factor returns and the market return. For example, the coefficient in the center-left panel is estimated using the regression:

$$R_{p^*i} = \alpha + \beta_m R_{m^*i} + \beta_s R_{si} + \varepsilon_i$$

Controlling for the market affects both coefficients: The size effect is halved, while the value effect is reduced by 30%. These changes are not surprising given the correlation between the size/value returns and the market return, which are reported in Table 8.2.

The final row shows the effect of controlling for both the returns of the market and the *other* factor. Only one model is estimated to produce the slopes in both panels:

$$R_{p^*i} = \alpha + \beta_m R_{m^*i} + \beta_s R_{si} + \beta_v R_{vi} + \varepsilon_i$$

The data points plotted are the residuals from regressions on the market return and the other factor return, and so control for the effects of variables that are not shown. For example, the points in the bottom-left panel depict the relationship between the computer industry portfolio returns and the size factor returns, controlling for the market and the value factor.

Adding the additional factor further reduces the magnitude of both coefficients. The reduction occurs because the factor returns are correlated, conditional on the market, and so the slopes reported in the center panels do not capture the distinct contribution of each factor.

The conditional correlation coefficient is the correlation between the residuals $\tilde{R}_{si}$ and $\tilde{R}_{vi}$, which are calculated from a regression of the factor return (i.e., R_{si} or R_{vi}) on the market return. For example:

$$\tilde{R}_{si} = R_{si} - \hat{\delta} - \hat{\gamma} R_{m^*i},$$

where $\hat{\delta}$ and $\hat{\gamma}$ are estimated using OLS.

These residuals have "purged" the effect of the market (i.e., the correlation computed using the residuals is conditional on the market). This conditional correlation of −22.8% is reported in the bottom panel of Table 8.2.

Table 8.2 **The Top Panel Contains the Estimated Correlation Matrix of the Factor Returns. The Bottom Panel Contains the Estimated Conditional Correlation Matrix for the Returns to the Size Portfolio and the Value Portfolio Controlling for the Market Return**

Correlation			
	Market	**Size**	**Value**
Market	1	0.216	−0.176
Size	0.216	1	−0.256
Value	−0.176	−0.256	1

Conditional Correlation		
	Size	**Value**
Size	1	−0.228
Value	−0.228	1

8.3 MEASURING MODEL FIT

Minimizing the squared residuals decomposes the total variation of the data of the dependent variable into two distinct components: one that captures the unexplained variation (due to the error in the model) and another that measures explained variation (which depends on both the estimated parameters and the variation in the explanatory variables).

The total variation in the dependent variable is called the total sum of squares (*TSS*), which is defined as the sum of the squared deviations of Y_i around the sample mean:

$$TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 \tag{8.6}$$

Each observation *i* of the dependent variable is decomposed into two components: the fitted value ($\hat{Y}_i$) and the estimated residual ($\hat{\varepsilon}_i$), so that:

$$Y_i = \hat{\alpha} + \hat{\beta}_1 X_{1i} + \ldots + \hat{\beta}_k X_{ki} + \hat{\varepsilon}_i = \hat{Y}_i + \hat{\varepsilon}_i \tag{8.7}$$

Transforming the data and the fitted values by subtracting the sample mean from both sides:

$$Y_i - \overline{Y} = \hat{Y}_i - \overline{Y} + \hat{\varepsilon}_i$$

ensures that both sides of the equation are mean zero. Finally, squaring and summing these values over all *i* from 1 to *n* yields

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y} + \hat{\varepsilon}_i)^2$$

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 + \sum_{j=1}^{n}\hat{\varepsilon}_j^2 \tag{8.8}$$

Note that the second line is missing the cross term between $(\hat{Y}_i - \overline{Y})$ and $\hat{\varepsilon}_i$. This cross term is 0 because the sample correlation between the fitted values $\hat{Y}_i$ and the estimated residuals $\hat{\varepsilon}_i$ is 0.[8]

[8] The first-order conditions, or normal equations, of the OLS minimization problem are $\sum_{i=1}^{n} X_{ji}\hat{\varepsilon}_i = 0$ for $j = 1, 2, \ldots k$, and $\sum_{i=1}^{n}\hat{\varepsilon}_i = 0$. These imply that the sample correlation is exactly 0 by construction, i.e., $\widehat{Cov}[X_j, \hat{\varepsilon}] = \widehat{Corr}[X_j, \hat{\varepsilon}] = 0$ for all *j*. The correlation is also 0 with $\hat{Y}_i$ because the fitted value is a linear function of the *X*'s, namely $\hat{Y}_i = \hat{\alpha} + \hat{\beta}_1 X_{1i} + \ldots + \hat{\beta}_k X_{ki}$.

The three components of this decomposition in the second line of equation (8.8) are written as:

$$TSS = ESS + RSS \tag{8.9}$$

are *TSS* (the total sum of squares), *ESS* (the explained sum of squares), and *RSS* (the residual sum of squares).

Normalizing this relationship by dividing both sides by *TSS* yields:

$$1 = \frac{ESS}{TSS} + \frac{RSS}{TSS}$$

This leads to the definition of R^2 in a model with any number of included explanatory variables:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \tag{8.10}$$

As an example, suppose the results of a model fit are as follows:

Observation Number	y	$\hat{y}$	$(y - \bar{y})^2$	$(\hat{y} - \bar{y})^2$
1	1.922	2.821	0.850	3.316
2	−2.367	−2.137	11.337	9.841
3	−4.868	−4.501	34.433	30.261
4	5.017	4.948	16.136	15.587
5	2.272	2.450	1.618	2.103
6	−0.596	0.062	2.547	0.880
7	2.329	2.721	1.766	2.962
8	4.295	3.894	10.857	8.375
9	−0.875	−1.238	3.516	5.009
10	1.259	1.538	0.067	0.289
11	0.442	0.425	0.311	0.331
12	−1.105	−1.283	4.431	5.212
13	3.524	1.116	6.371	0.013
14	0.971	1.932	0.001	0.869
15	2.377	2.259	1.896	1.585
16	5.445	6.163	19.758	26.657
17	−0.331	0.044	1.772	0.914
18	−3.638	−2.010	21.511	9.060
19	2.949	0.375	3.799	0.391
20	0.978	0.421	0.000	0.335
Average	1.000			
		SUM	142.977	123.988

Accordingly:

$$R^2 = \frac{ESS}{TSS} = \frac{123.988}{142.977} = 86.7\%$$

It can then be deduced that:

$$\frac{RSS}{TSS} = 1 - 0.867 = 13.3\%$$

and:

$$RSS = 0.133 \times 142.977 = 19.02$$

R^2 measures the percentage of the variation in the data that can be explained by the model. This is equivalently measured by the ratio of the explained variation to the total variation (which is equal to one minus the ratio of the residual variation to the total variation). Since OLS estimates parameters by finding the values that minimize the *RSS*, the OLS estimator also maximizes R^2.

As explained in Chapter 7, R^2 is also defined as the squared correlation between the dependent variable and the explanatory variable in a model with a single explanatory variable. In such a model, the fitted value is a linear transformation of the explanatory variable (i.e., $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$) and so the correlation between Y_i and X_i is the same as the correlation between Y_i and $\hat{Y}_i$ (up to the sign, which changes if $\hat{\beta}$ is negative).

When a model has multiple explanatory variables, R^2 is a complicated function of the correlations among the explanatory variables and those between the explanatory variables and the dependent variable. However, R^2 in a model with multiple regressors *is* the squared correlation between Y_i and the fitted value $\hat{Y}_i$:

$$R^2 = \widehat{\text{Corr}}^2[Y, \hat{Y}]$$

This definition of R^2 provides another interpretation of the OLS estimator: the regression coefficients $\hat{\beta}_1, \ldots, \hat{\beta}_k$ are chosen to produce the linear combination of $X_1, \ldots, X_k$ that maximizes the correlation with Y.

A model that is completely incapable of explaining the observed data has an R^2 of 0 (because all variation is in the residuals). A model that perfectly explains the data (so that all residuals are 0) has an R^2 of 1. All other models must produce values that fall between these two bounds so that R^2 is never negative and always less than 1.

While R^2 is a useful method to assess model fit due to its simplicity and ease of interpretation, it has three important limitations.

1. Adding a new variable to the model always increases the R^2 even if the new variable is almost totally irrelevant for explaining Y. For example, if the original model is $Y_i = \alpha + \beta_1 X_{1i} + \varepsilon_i$ and the expanded model $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$ is estimated, then the R^2 of the expanded model *must* be greater than or equal to the R^2

of the original model. This is because the expanded model always has the same *TSS* and nearly always has a smaller *RSS*, resulting in a higher R^2. The only situation where adding a variable does not increase R^2 is if $\beta_2 = 0$. In that case, the *RSS* remains the same (as does the R^2).

2. R^2 cannot be compared across models with different dependent variables. For example, when Y_i is always positive, it is not possible to compare the R^2 of a model in levels (Y_i) and logs ($\ln Y_i$). It is also not possible to compare the R^2 for two models that are logically equivalent (in the sense that both the fit of the model as measured by *RSS* and predictions from the models are identical). The second case can occur when the dependent variable is transformed by adding or subtracting one or more of the explanatory variables. For example, consider:

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

and:

$$Y_i - X_i = \delta + \gamma X_i + \eta_i$$

OLS estimates of the slope in these two specifications of the same model have the relationship that $\hat{\gamma} + 1 = \hat{\beta}$. While these two models have the same *RSS*, they have different *TSS* and therefore different R^2. These models are therefore statistically identical yet cannot be compared using R^2.

3. There is no such thing as a "good" value for R^2. Whether a model provides a good description of the data depends on the nature of the data on the dependent variable. For example, an R^2 of 5% would be implausibly high for a model that predicts the one-day ahead return on a liquid equity index futures contract using the current value of explanatory variables. On the other hand, an R^2 less than 70% would be quite low for a model for predicting the returns on a well-diversified large-cap portfolio using the contemporaneous return on the equity market (i.e., CAPM).

The first limitation is addressed (in a limited way) by the adjusted R^2, which is written as $\bar{R}^2$. The adjusted R^2 modifies the standard R^2 to account for the degrees of freedom used when estimating model parameters, so that:

$$\begin{aligned}\bar{R}^2 &= 1 - \frac{RSS/(n-k-1)}{TSS/(n-1)} \\ &= 1 - \frac{n-1}{n-k-1}(1 - R^2), \qquad (8.11)\end{aligned}$$

where n is the number of observations in the sample and k is the number of explanatory variables included in the model (not including the constant term α).

Adjusted R^2 can also be expressed as:

$$\bar{R}^2 = 1 - \xi\frac{RSS}{TSS},$$

where the adjustment factor $\xi = (n-1)/(n-k-1)$. Note that ξ must be greater than 1 because the denominator is less than the numerator.

As explained previously, adding an additional explanatory variable usually increases R^2 and can never decrease it. On the other hand, including additional explanatory variables (i.e., increasing k) always increases ξ. The adjusted R^2 captures the tradeoff between increasing ξ and decreasing *RSS* as one considers larger models. If a model with additional explanatory variables produces a negligible decrease in the *RSS* when compared to a base model, then the loss of a degree of freedom produces a smaller $\bar{R}^2$.

The adjustment to the R^2 may produce negative values if a model produces an exceptionally poor fit. In most financial data applications, n is relatively large and so the loss of a degree of freedom has little effect on $\bar{R}^2$. In large samples, the adjustment term ξ is very small and so $\bar{R}^2$ tends to increase even when an additional variable has little explanatory power.

8.4 TESTING PARAMETERS IN REGRESSION MODELS

When the assumptions in Section 8.1 are satisfied, the parameter estimators $\hat{\alpha}$ and $\hat{\beta}_j$ follow a CLT. Thus, testing a hypothesis about a single coefficient in a model with multiple regressors is identical to testing in a model with a single explanatory variable. Tests of the null hypothesis H_0: $\beta_j = \beta_{j0}$ are implemented using a *t*-test:

$$\frac{\hat{\beta}_j - \beta_{j0}}{\widehat{\text{s.e.}}(\hat{\beta}_j)}, \qquad (8.12)$$

where $\widehat{\text{s.e.}}(\hat{\beta}_j)$ is the estimated standard error of $\hat{\beta}_j$. While the precise formula for the standard error is complicated in the *k*-variable model, it is a routine output in all statistical software packages.

However, the *t*-test is *not* directly applicable when testing hypotheses that involve more than one parameter, because the parameter estimators can be correlated.[9] This correlation complicates extending the univariate *t*-test to tests of multiple parameters.

Instead, the more common choice is an alternative called the *F*-test. This type of test compares the fit of the model (measured using the *RSS*) when the null hypothesis is true relative to the fit of the model without the restriction on the parameters assumed by the null.

[9] Specifically, the parameter estimators will be correlated if the explanatory variables are correlated.

Implementing an F-test requires estimating two models. The first is the full model that is to be tested. This model is called the unrestricted model (because it is specified before any restrictions are applied) and has an RSS denoted by RSS_U. The second model, called the restricted model, imposes the null hypothesis on the unrestricted model and its RSS is denoted RSS_R. The F-test compares the fit of these two models:

$$F = \frac{(RSS_R - RSS_U)/q}{RSS_U/(n - k_U - 1)}, \qquad (8.13)$$

where q is the number of restrictions imposed on the unrestricted model to produce the restricted model and k_U is the number of explanatory variables in the unrestricted model. An F-test has an $F_{q,n-k_U-1}$ distribution.

F-tests can be equivalently expressed in terms of the R^2 from the restricted and unrestricted models. Using this alternative parameterization:

$$F = \frac{(R_U^2 - R_R^2)/q}{(1 - R_U^2)/(n - k_U - 1)} \qquad (8.14)$$

This expression is numerically identical to Equation (8.13).[10]

If the restriction imposed by the null hypothesis does not meaningfully alter the fit of the model, then the two RSS measures are similar, and the test statistic is small. On the other hand, if the unrestricted model fits the data significantly better than the restricted model, then the RSS from the two models should differ by a large amount so that the value of the F-test statistic is large. A large test statistic indicates that the unrestricted model provides a superior fit and so the null hypothesis is rejected.

Implementing an F-test requires imposing the null hypothesis on the model and then estimating the restricted model using OLS. For example, consider a test of whether CAPM, which only includes the market return as a factor, provides as good a fit as a multi-factor model that additionally includes the size and value factors.

The unrestricted model includes all three explanatory variables, so that:

$$R_{p^*i} = \alpha + \beta_m R_{m^*i} + \beta_s R_{si} + \beta_v R_{vi} + \varepsilon_i,$$

where m indicates the market (i.e., so that R_{m^*i} is the return to the market factor above the risk-free-rate), s indicates size, and v indicates value. The null hypothesis is then:

$$H_0: \beta_s = \beta_v = 0$$

[10] Transforming between the two forms relies on two identities: $1 - R_U^2 = RSS_U/TSS$ and $R_U^2 - R_R^2 = (1 - R_R^2) - (1 - R_U^2)$ so that $R_U^2 - R_R^2 = (RSS_R - RSS_U)/TSS$.

The alternative hypothesis is that at least one of parameters is not equal to zero:

$$H_1: \beta_s \neq 0 \text{ or } \beta_v \neq 0$$

so that the null should be rejected if at least one of the coefficients is different from zero.

Imposing the null hypothesis requires replacing the parameters with their assumed value if the null is true. Imposing the null hypothesis on the unrestricted model produces the restricted model:

$$\begin{aligned} R_{p^*i} &= \alpha + \beta_m R_{m^*i} + 0 \times R_{si} + 0 \times R_{vi} + \varepsilon_i \\ &= \alpha + \beta_m R_{m^*i} + \varepsilon_i, \end{aligned}$$

which is the CAPM.

In this hypothesis test, two coefficients are restricted to specific values and so the number of restrictions is $q = 2$. The F-test is then computed by estimating both regressions, storing the two RSS values, and then computing:

$$F = \frac{(RSS_R - RSS_U)/2}{RSS_U/(n - 4)}$$

Finally, if the test statistic F is larger than the critical value of an $F_{2,n-4}$ distribution using a size of α (e.g., 5%), then the null is rejected. If the test statistic is smaller than the critical value, then the null hypothesis is not rejected, and it is concluded that CAPM appears to be adequate in explaining the returns to the portfolio.

Table 8.3 contains the value of the F-test statistics of this null hypothesis applied to the industry portfolios. The p-value of the test statistic, which measures the probability of observing the value of the test statistic if the null is true, is reported below each test statistic. The p-value is computed as $1 - C(F)$, where C is the CDF of an $F_{2,n-4}$ distribution. A p-value of less than 5% indicates that the null is rejected when using a 5% test, which is the case for all portfolios except the electrical equipment and retail industry portfolios.

The second column of numbers in Table 8.3 tests the joint null hypothesis that the parameters for the size and value factors are both zero and that the parameter on the market portfolio is 1 (i.e., that the portfolio has unit sensitivity to movements in the stock market as a whole). This null is very strongly rejected for all portfolios except for Shipping Containers, where it is rejected if a 10% significance level is used, and the Retail sector.

Multivariate Confidence Intervals

Recall that for a single parameter, a $1 - \alpha$ confidence interval defines the values where a test with size α is not rejected. The same idea can be applied to an F-test to define a region where the null hypothesis is not rejected.

Table 8.3 **The Three Columns Report the Test Statistic Value Using F-Tests for the Nulls $H_0: \beta_S = \beta_v = 0$ (Left), $H_0: \beta_m = 1$ and $\beta_s + \beta_v = 0$ (Center), and the Final Column Reports the F-Statistic of the Regression That Tests $H_0: \beta_m = \beta_s = \beta_v = 0$ (Right). The Values Below the Parameters in Parentheses Are the p-values of the Test Statistics, Which are Computed as $1 - C(F)$, Where C Is the CDF of the Distribution of the Test Statistic**

	$H_o: \beta_s = \beta_v = 0$	$H_o: \beta_m = 1$ and $\beta_s + \beta_v = 0$	F-statistic
Banking	135.3 (0.000)	51.8 (0.000)	575.7 (0.000)
Shipping Containers	6.6 (0.002)	2.5 (0.087)	176.9 (0.000)
Chemicals	32.1 (0.000)	14.5 (0.000)	372.8 (0.000)
Electrical Equipment	2.2 (0.111)	19.3 (0.000)	450.5 (0.000)
Computers	40.8 (0.000)	17.8 (0.000)	344.3 (0.000)
Consumer Goods	9.2 (0.000)	33.6 (0.000)	143.4 (0.000)
Retail	0.3 (0.763)	1.7 (0.182)	282.9 (0.000)

Figure 8.2 contains four examples of bivariate confidence regions.[11] These regions are multidimensional ellipses that account for the variances of each parameter and the correlation between them. The confidence regions in Figure 8.2 are all constructed from the model:

$$Y_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_i,$$

where $\beta_1 = \beta_2 = 1$. The errors are standard normal random variables, and the explanatory variables are bivariate normal:

$$X_i \overset{iid}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{11} & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_{22} \end{bmatrix}\right)$$

The baseline specification uses a bivariate standard normal so that $\sigma_{11} = \sigma_{22} = 1$ and $\rho = 0$. The confidence region in the baseline specification with uncorrelated regressors is shown in the top left panel. This region is spherical, which reflects the lack of correlation between the estimated estimators $\hat{\beta}_1$ and $\hat{\beta}_2$.

[11] Confidence intervals are used with a single parameter. "Confidence region" is the preferred term when constructing the regions where the null is not rejected for more than one parameter.

The top-right panel is generated from a model where the correlation between the explanatory variables is $\rho = 50\%$. The positive correlation between the variables creates a negative correlation between the estimated parameters. When the explanatory variables are positively related and $\hat{\beta}_1$ is overestimated (i.e., so $\hat{\beta}_1 > \beta_1$), then $\hat{\beta}_2$ tends to underestimate β_2 to compensate. The negative correlation mitigates the common component in the two correlated explanatory variables.

The bottom-left panel shows the case where the correlation is large and negative (−80%). The confidence interval is upward sloping, which reflects the positive correlation between parameter estimators. Because the variables have a negative correlation, $\hat{\beta}_2$ will tend to overestimate β_2 to compensate if $\hat{\beta}_1$ is overestimated. The compensation improves the fit because if X_1 is above its mean, X_2 will tend to be below its mean due to the negative correlation.

The bottom-right shows a different scenario where $\rho = 50\%$ and $\sigma_{11} = 0.5$. The length of the region for each variable is proportional to the inverse of its variance, therefore, halving the variance of X_1 doubles the width of the region along the x-axis relative to the upper-right panel.

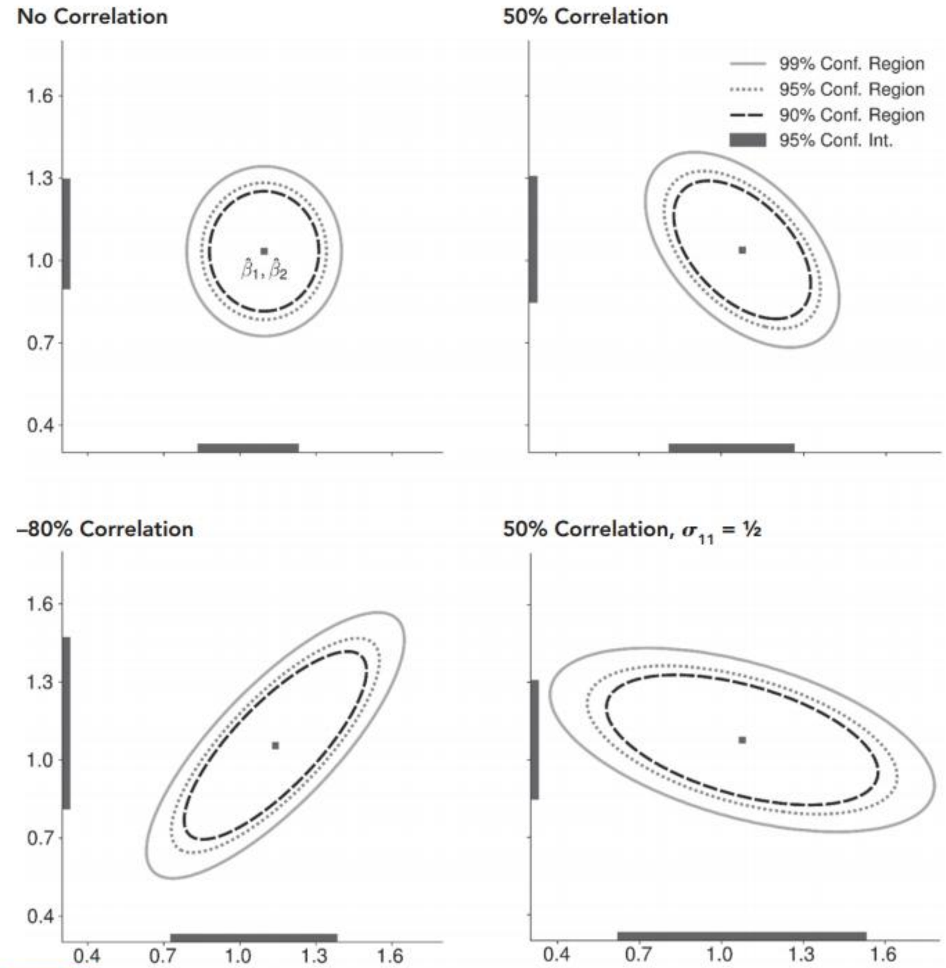


Figure 8.2 Illustration of four bivariate confidence intervals. Each confidence interval is centered on the estimated parameters $\hat{\beta}_1$ (x-axis) and $\hat{\beta}_2$ (y-axis). The top-left panel shows the confidence interval when the two explanatory variables in the model are uncorrelated. The top-right panel shows the confidence interval when the variables have identical variances but a correlation of 50%. The bottom-left panel shows the confidence interval when the correlation is -80%. The bottom-right shows the confidence interval when the correlation is 50% but the variance of $\hat{\beta}_1$ is twice the variance of $\hat{\beta}_2$. The dark bars along the axes show individual (univariate) 95% confidence intervals for $\hat{\beta}_1$ and $\hat{\beta}_2$.

CONFLICTING *F*- AND *t*-TEST STATISTICS

Suppose that a model with two explanatory variables is to be fitted:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i.$$

The significance of this relationship can be examined by:

- Using an *F*-test of the null $H_0: \beta_1 = \beta_2 = 0$, or
- Testing each of the two coefficients separately using *t*-tests of the null $H_0: \beta_j = 0$ for $j = 1, 2$.

Under most circumstances, the conclusions drawn from these two methods agree (i.e., both testing methods should indicate that at least one of the parameters is different from zero, or both should fail to reject their respective null hypotheses).

Disagreement between the two types of tests is driven by one of two issues. When the *F*-test rejects the null but the *t*-tests do not, then it is likely that the data are multicollinear. The *F*-test tells us that the model is providing a useful fit of the data. The *t*-tests indicate that the fit of the model cannot be uniquely attributed to either of the two explanatory variables. For example, if X_1 and X_2 are extremely highly correlated (say 98%+), then the model fit from $\beta_1 = \beta_2 = 1$ is similar to any other model fit where the coefficients sum to 1 (e.g., $\beta_1 = 1.5$ and $\beta_2 = 0.5$). In models with substantial multicollinearity, it is simple to detect an effect of the two but difficult to uniquely attribute it to either.

On the other hand, if the *F*-test fails to reject but one of the *t*-tests *does*, then this indicates that one variable has a very small effect. If X_1 and X_2 are uncorrelated, it can be shown that:

$$F \approx \frac{T_1^2 + T_2^2}{2},$$

where *F* is the *F*-test statistic and T_i is the *t* test statistic for $\hat{\beta}_i$. If variable 2 has no effect on *Y* (i.e., $T_2 \approx 0$), then $F \approx T_1^2/2$. The *F*-test will only indicate significance if T_1 is moderately large. For example, when $n = 250$, then T_1^2 would need to be larger than 6.06 to reject the null if $T_2 = 0$, which would require $|T_1| > 2.46$. Treated as a single test, any value to $|T_1| > 1.96$ indicates that the null $H_0: \beta_1 = 0$ should be rejected, and so when $1.96 \leq |T_1| < 2.46$, the tests disagree.

The *F*-statistic of a Regression

The *F*-statistic of a regression is the multivariate generalization of the *t*-statistic of a coefficient. It is an *F*-test of the null that all explanatory variables have 0 coefficients:

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

The alternative hypothesis is:

$$H_1: \beta_j \neq 0$$

for some *j* in 1, 2 . . . , *k*. The test should reject whenever one or more of the regression coefficients are nonzero.

The null hypothesis does not assume any value for the intercept, α, and so the restricted model only includes an intercept:

$$Y_i = \alpha + \varepsilon_i$$

The test statistic:

$$F = \frac{(TSS - RSS_U)/k}{RSS_U/(n - k - 1)} \sim F_{k,n-k-1} \qquad (8.15)$$

depends on the *TSS*, because $TSS = RSS_R$ when the restricted model has no explanatory variables. Also, note that $q = k$ because every coefficient is being tested.

The final column of Table 8.3 reports the *F*-statistic for the three-factor models of the industry portfolio returns in Table 8.1. The portfolio returns are strongly related to the market, and so it is not surprising that all *F*-statistics indicate that the null of all coefficients being zero is rejected.

8.5 SUMMARY

This chapter describes the *k*-variable linear regression model, which extends the single explanatory variable model to account for additional sources of variation. All key concepts from the simple specification carry over to the general model.

However, the presence of multiple explanatory variables changes how coefficients are interpreted. In most cases, the coefficients are interpreted as the effect of a small change in one explanatory variable holding all others constant. This means that the OLS estimators measure the unique contribution of each explanatory variable using a component of the variable that is uncorrelated with the other included variables.

Model fit is assessed using R^2, which measures the ratio of the variation explained by the model to the total variation in the data. While intuitive, this measure suffers from some important limitations: It never decreases when an additional variable is added to a model and is it not interpretable when the dependent variable changes. The adjusted R^2 ($\bar{R}^2$) partially addresses the first of these concerns.

The same variation measures that appear in R^2 are used to test model parameters with *F*-tests. This test statistic measures the difference in the fit between a restricted model (which imposes the null hypothesis) and the unrestricted model. If the restriction is valid, the fits of the two models should be similar. When

the restriction is invalid, the restricted model should fit the data considerably worse than the unrestricted model so that the test statistic is large, and the null should be rejected. Most statistical software packages report the *F*-statistic of the regression, which tests the null hypothesis that all of the explanatory variables have zero coefficients.

8.6 APPENDIX

Tables

These tables contain the values that are used in the example that runs through the chapter.

First-Step Regressions

This table contains the annual returns from 1999 until 2018 for three series:

1. The excess return on the shipping container industry, R_{p^*};
2. The excess return on the market portfolio, R_{m^*}; and
3. The return of the value factor, R_v.

The columns labeled using variable names starting with *E* are the squares and cross-products of the demeaned data. For example, $E_P = R_{p^*} - \overline{R_{p^*}}$ is the demeaned excess return on the shipping container return, and $E_p^2 = (R_{p^*} - \overline{R_{p^*}})^2$. The columns labeled with two *E* variables are cross-products, so that $E_pE_m = (R_{p^*} - \overline{R_{p^*}})(R_{m^*} - \overline{R_{m^*}})$.

Year	R_{p^*}	R_{m^*}	R_v	E_p^2	E_m^2	E_v^2	E_pE_m	E_vE_m
1999	−5.46	19.7	−24.8	197.2	186.9	715.2	−192.0	−365.6
2000	−37.7	−16.7	39.2	2142.0	516.5	1388.1	1051.9	−846.8
2001	22.9	−14.7	15.8	205.0	429.6	192.0	−296.8	−287.2
2002	14.4	−22.4	8.6	33.8	808.1	44.3	−165.4	−189.3
2003	11.9	30.5	3.79	11.0	598.9	3.4	81.2	45.2
2004	36.5	10.6	7.1	779.4	20.9	26.6	127.7	23.6
2005	−2.32	3	7.86	118.9	9.2	35.0	33.0	−17.9
2006	12.1	10.2	12.5	12.4	17.4	111.5	14.7	44.1
2007	24	0.976	−13.6	237.7	25.5	241.6	−77.9	78.5
2008	−30.4	−37.8	1.71	1519.6	1920.8	0.1	1708.5	10.2
2009	34.8	28.2	−4.28	687.4	491.6	38.7	581.3	−138.0
2010	17.7	17.4	−3.85	83.1	129.3	33.6	103.7	−65.9
2011	−5.87	0.415	−8.26	208.9	31.5	104.1	81.1	57.3
2012	18.1	16.3	8.36	90.6	105.5	41.2	97.8	65.9
2013	36.7	35.2	1.31	790.6	851.1	0.4	820.3	−18.5
2014	12.8	11.7	−1.92	17.8	32.2	14.9	23.9	−21.9
2015	−5.31	0.06	−9.87	193.0	35.6	139.5	82.9	70.5
2016	19.2	13.3	20.7	112.7	52.9	351.8	77.2	136.4
2017	16.9	21.4	−11.2	69.2	236.3	172.7	127.9	−202.0
2018	−19.3	−6.81	−10.3	777.4	164.8	149.9	357.9	157.2
Sum	171.6	120.5	38.9	8287.7	6664.8	3804.6	4638.9	−1464.2
Mean	8.58	6.03	1.94	414.39	333.24	190.23	231.94	−73.21

Second-Step Regressions

This table contains the values of the shipping container industry return residuals ($\tilde{R}_{p^*}$) and the value factor ($\tilde{R}_v$) from a regression on a constant and the excess return on the market. These are, by construction, mean zero and uncorrelated with the excess return on the market. The final three columns contain the squares of these residuals and their cross-products.

Year	$\tilde{R}_{p^*}$	$\tilde{R}_v$	$\tilde{R}_{p^*}^2$	$\tilde{R}_v^2$	$\tilde{R}_{p^*}\tilde{R}_v$
1999	−23.6	−23.74	555	564	559.3
2000	−30.5	32.3	928	1041	−982.9
2001	28.7	9.3	826	87	267.4
2002	25.6	0.4	656	0	10.6
2003	−13.7	7.2	188	52	−99.1
2004	24.7	6.2	612	38	152.4
2005	−8.8	5.3	77	28	−46.2
2006	0.6	11.5	0	132	7.0
2007	18.9	−16.7	358	277	−315.3
2008	−8.5	−9.9	72	97	83.6
2009	10.8	−1.4	116	2	−14.6
2010	1.2	−3.3	1	11	−4.0
2011	−10.5	−11.4	111	131	120.6
2012	2.4	8.7	6	75	20.5
2013	7.8	5.8	61	33	45.1
2014	0.3	−2.6	0	7	−0.7
2015	−9.7	−13.1	95	172	127.8
2016	5.6	20.4	31	414	113.1
2017	−2.4	−9.8	6	95	23.3
2018	−18.9	−15.1	359	227	285.4
Sum	0	0	5059	3483	353
Mean	0	0	253	174	17.7

The following questions are intended to help candidates understand the material. They are *not* actual FRM exam questions.

QUESTIONS

Short Concept Questions

8.1 How does the correlation between pairs of explanatory variables affect the interpretation of the coefficients in a regression with multiple explanatory variables?

8.2 Both R^2 and $\bar{R}^2$ can only increase when adding a new variable. True or false?

8.3 When is $\bar{R}^2$ less than 0? Can this measure of fit be larger than 1?

8.4 The value of an F-test can be negative. True or false?

8.5 You estimate a regression model $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$. Using the F-stat of the model, you reject the null $H_0: \beta_1 = \beta_2 = 0$ but fail to reject either of the nulls $H_0: \beta_1 = 0$ or $H_0: \beta_2 = 0$ using the t-stat of the coefficient. Which values of $\rho = \text{Corr}[X_1, X_2]$ make this scenario more likely?

8.6 Suppose you fit a model where both t-statistics are exactly +1.0. Could an F-statistic lead to rejection in this circumstance? Use a diagram to help explain why or why not.

Practice Questions

8.7 Construct 95% confidence intervals and p-values for the coefficients in the shipping containers industry portfolio returns.

8.8 The models were estimated on monthly returns. What is the estimate of the annualized α for the beer and liquor industry portfolio? What is the t-statistic for the annualized α?

8.9 Suppose that Y_i and X_i follow a bivariate normal so that:

$$\begin{bmatrix} Y_i \\ X_i \end{bmatrix} \overset{iid}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.9 \\ 0.9 & 3 \end{bmatrix}\right)$$

What is the R^2 in a model that regresses Y on X? What would the R^2 be in a model that regressed X on Y?

8.10 A model is estimated using daily data from the S&P 500 from 1977 until 2017 which includes five day-of-the-week dummies ($n = 10{,}087$). The R^2 from this regression is 0.000599. Is there evidence that the mean varies with the day of the week?

8.11 Using the data below:

Trial Number	y	x_1	x_2
1	−5.76	−3.48	−1.37
2	0.03	−0.02	−0.62
3	−0.25	−0.5	−1.07
4	−2.72	−0.18	−1.01
5	−3.08	−0.82	0.39
6	−7.1	−2.08	1.39
7	−4.1	−1.06	0.75
8	0.14	0.02	−0.63
9	−6.13	−1.66	1.31
10	0.74	0.68	−0.15

a. Apply OLS linear regression to find the parameters for the following equation:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

b. What is the R^2, $\bar{R}^2$, and F statistic of this regression?

ANSWERS

Short Concept Questions

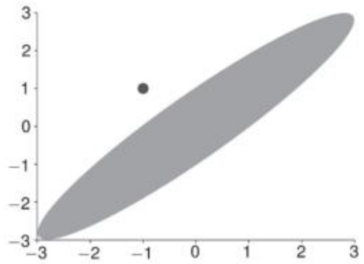
8.1 Regression coefficients are interpreted as holding all other variables constant, and so the correlation between the regressors is immaterial to the interpretation of the coefficients.

8.2 False. R^2 will always increase. $\bar{R}^2$ can decrease if the degree-of-freedom adjusted residual variance estimator, $\frac{\sum \hat{\varepsilon}_i^2}{n-k}$, does not decrease when adding a new regressor. This can happen because k is the number of regressors in the model and so if the numerator does not change, adding an additional regressor will increase k and so decrease the denominator.

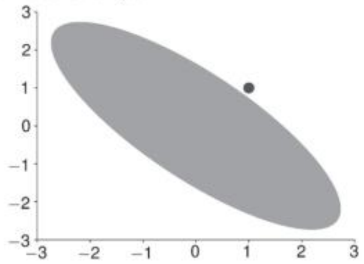
8.3 $\bar{R}^2$ is always less than R^2 and so cannot be larger than 1. It can be smaller than 0 if the model explains virtually nothing and the degree-of-freedom loss is sufficient to push its value negative.

8.4 False. An *F*-test measures the reduction in the sum of squared residuals that occurs in the expanded model and depends on $SSR_U - SSR_R$. This value is always positive because the unrestricted model must fit at least as well the restricted model.

8.5 This is most likely to occur when the regressors are highly correlated. If the regressors are positively correlated, then the parameter estimators of the coefficients will be negatively correlated. If both values are positive, this would lead to rejection by the *F*-test. Similarly, if the regressors were negatively correlated, then the estimators are positively correlated and the *F* will reject if one *t* is positive and the other is negative. The figure below shows the case for positively correlated regressors. The shaded region is the area where the *F* would fail to reject. The *t*-stats are outside this area even though neither is individually significant.



8.6 Yes. When the regressors are *very* positively correlated, then this can happen. The *t*-stats are small because the variables are highly co-linear, so that the variation in the left-hand-side variable cannot be uniquely attributed to either. However, the *F*-stat is a joint test and so if there is an effect in at least one, the *F* can reject even if the *t*-stats do not. See the image below that shows the region where the *F* would fail to reject.



Solved Problems

8.7

Coefficient	Estimate	t-stat	Std. Err.	Conf. Int.	p-value
β_m	1.011	18.63	0.054	[0.905, 1.117]	0.000
β_s	−0.012	−0.16	0.075	[−0.159, 0.135]	1.128
β_v	0.276	3.51	0.079	[0.121, 0.431]	0.002

The confidence interval is $\hat{\beta} \pm 1.96 \times se$. The p-value is $2(1 - \Phi(|t|))$ where $\Phi(\cdot)$ is the standard normal CDF.

8.8 The annualized α is $12\hat{\alpha} = 12 \times 0.517 = 6.204\%$. The *t*-stat is unaffected because the standard error is also scaled by 12.

8.9 Because this is a regression with a single explanatory variable, the R^2 is the squared correlation. The correlation is $0.9/\sqrt{3} = .519$ and so the $R^2 = 0.27$. The R^2 is

the same in both regressions because it only depends on the correlation between the variables.

8.10 The model estimated is
$Y_i = \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 + \beta_4 D_4 + \beta_4 D_5 + \varepsilon_i$, where D_i is a dummy that takes the value 1 if the index of the weekday is i (e.g., Monday = 1, Tuesday = 2, ...).
The restriction is that $H_0{:}\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5$ so there this is no day-of-the-week effect.
This model can be equivalently written as
$Y_i = \mu + \delta_2 D_2 + \delta_3 D_3 + \delta_4 D_4 + \delta_4 D_5 + \varepsilon_i$, therefore, here the null is $H_0{:}\delta_2 = \delta_3 = \delta_4 = \delta_5 = 0$. In the two models, $\mu = \beta_1$, and $\mu + \delta_i = \beta_i$. The second form of the model is a more standard null for an F-stat.

The F-stat of the regression is

$$\frac{R^2 - 0 / 4}{1 - R^2 / n-5} = \frac{\left(0.000599 / 4\right)}{1 - 0.000599 / (10087 - 5)} = 1.51.$$ The distribution is an $F_{4,10082}$ and the critical value using a 5% size is 2.37. The test statistic is less than the critical value, therefore, the null that all effects are 0 is not rejected.

8.11 **a.** Start by setting up the basic calculations:

Trial Number	y	x_1	x_2
1	−5.76	−3.48	−1.37
2	0.03	−0.02	−0.62
3	−0.25	−0.5	−1.07
4	−2.72	−0.18	−1.01
5	−3.08	−0.82	0.39
6	−7.1	−2.08	1.39
7	−4.1	−1.06	0.75
8	0.14	0.02	−0.63
9	−6.13	−1.66	1.31
10	0.74	0.68	−0.15
Average	−2.823	−0.910	−0.101

Trial Number	$y - \bar{y}$	$x_1 - \overline{x_1}$	$x_2 - \overline{x_2}$
1	−2.937	−2.570	−1.269
2	2.853	0.890	−0.519
3	2.573	0.410	−0.969
4	0.103	0.730	−0.909
5	−0.257	0.090	0.491
6	−4.277	−1.170	1.491
7	−1.277	−0.150	0.851
8	2.963	0.930	−0.529
9	−3.307	−0.750	1.411
10	3.563	1.590	−0.049
Average			

Trial Number	$(y - \bar{y})^2$	$(x_1 - \overline{x_1})^2$	$(x_2 - \overline{x_2})^2$	$(x_1 - \overline{x_1})(y - \bar{y})$	$(x_2 - \overline{x_2})(y - \bar{y})$	$(x_1 - \overline{x_1})(x_2 - \overline{x_2})$
1	8.626	6.605	1.610	7.548	3.727	3.261
2	8.140	0.792	0.269	2.539	−1.481	−0.462
3	6.620	0.168	0.939	1.055	−2.493	−0.397
4	0.011	0.533	0.826	0.075	−0.094	−0.664
5	0.066	0.008	0.241	−0.023	−0.126	0.044
6	18.293	1.369	2.223	5.004	−6.377	−1.744
7	1.631	0.023	0.724	0.192	−1.087	−0.128
8	8.779	0.865	0.280	2.756	−1.567	−0.492
9	10.936	0.563	1.991	2.480	−4.666	−1.058
10	12.695	2.528	0.002	5.665	−0.175	−0.078
Average	7.580	1.345	0.911	2.729	−1.434	−0.172

This gives

$\mu_Y = -2.823 \quad \mu_{X_1} = -0.91 \quad \mu_{X_2} = -0.101$

$\sigma_Y^2 = 7.58 \quad \sigma_{X_1}^2 = 1.345 \quad \sigma_{X_2}^2 = 0.911$

$\sigma_{X_1X_2} = -0.172 \quad \sigma_Y\sigma_{X_1} = 2.729 \quad \sigma_Y\sigma_{X_2} = -1.434$

First regressing on the last independent variable:

$X_{1i} = \hat{\delta}_0 + \hat{\delta}_1 X_{2i} + \bar{X}_{1i}$

$\hat{\delta}_1 = \frac{\sigma_{X_1X_2}}{\sigma_{X_2}^2} = \frac{-0.172}{0.911} = 0.189$

$\hat{\delta}_0 = \mu_{X_1} - \hat{\delta}_1\mu_{X_2} = -0.91 - 0.189{*}(-0.101) = -0.929$

$Y_i = \hat{\gamma}_0 + \hat{\gamma}_1 X_{2i} + \bar{Y}_i$

$\hat{\gamma}_1 = \frac{\sigma_{YX_2}}{\sigma_{X_2}^2} = \frac{-1.434}{0.911} = -1.574$

$\hat{\gamma}_0 = \mu_Y - \hat{\gamma}_1\mu_{X_2} = -2.823 - (-1.574{*}-0.101) = -2.982$

Trial Number	$\bar{y}$	$\bar{x}_1$	$\bar{y}\bar{x}_1$	$\bar{x}_1^2$
1	−4.934	−2.810	13.865	7.896
2	2.036	0.792	1.612	0.627
3	1.048	0.227	0.238	0.051
4	−1.328	0.558	−0.741	0.311
5	0.516	0.183	0.094	0.033
6	−1.930	−0.888	1.715	0.789
7	0.063	0.011	0.001	0.000
8	2.130	0.830	1.768	0.689
9	−1.086	−0.483	0.525	0.234
10	3.486	1.581	5.510	2.498
Average			2.459	1.313

So:

$$\beta_1 = \frac{Cov(\bar{Y}, \bar{X}_1)}{\sigma_{\bar{X}_1}^2} = \frac{2.459}{1.313} = 1.873$$

Repeating the process for X_2:

$X_{2i} = \hat{\zeta}_0 + \hat{\zeta}_1 X_{1i} + \bar{X}_{2i}$

$\hat{\zeta}_1 = \frac{\sigma_{X_1X_2}}{\sigma_{X_1}^2} = \frac{-0.172}{1.345} = -0.128$

$\hat{\zeta}_0 = \mu_{X_2} - \hat{\zeta}_1\mu_{X_1} = -0.101 - (-0.128) - (-0.91) = -0.217$

$Y_i = \hat{\eta}_0 + \hat{\eta}_1 X_{1i} + \bar{K}_i$

$\hat{\eta}_1 = \frac{\sigma_{YX_1}}{\sigma_{X_1}^2} = \frac{2.729}{1.345} = 2.029$

$\hat{\eta}_0 = \mu_Y - \hat{\eta}_1\mu_{X_1} = -2.823 - (2.029)(-0.91) = -0.976$

So:

Trial Number	$\bar{k}$	$\bar{x}_2$	$\bar{k}\bar{x}_2$	$\bar{x}_2^2$
1	2.274	−1.598	−3.635	2.555
2	1.047	−0.406	−0.425	0.164
3	1.741	−0.917	−1.596	0.841
4	−1.378	−0.816	1.125	0.666
5	−0.440	0.502	−0.221	0.252
6	−1.905	1.341	−2.554	1.798
7	−0.974	0.831	−0.809	0.691
8	1.076	−0.410	−0.442	0.168
9	−1.787	1.315	−2.349	1.728
10	0.338	0.154	0.052	0.024
Average			−1.085	0.889

and

$$Cov(\bar{K}, \bar{X}_2) = -1.085 \quad \sigma_{\bar{X}_2}^2 = 0.889$$

$$\beta_2 = \frac{-1.085}{0.889} = -1.221$$

Tying things together:

$$Y_i - \beta_1 X_{1i} - \beta_2 X_{2i} = \alpha + \varepsilon_i = -1.873X_{1i} + 1.221X_{2i}$$

Trial Number	$\alpha + \epsilon_i$	ϵ_i	ϵ_i^2
1	−0.913	0.328	0.108
2	−0.689	0.553	0.306
3	−0.619	0.623	0.388
4	−3.615	−2.373	5.632
5	−1.068	0.173	0.030
6	−1.508	−0.267	0.071
7	−1.200	0.042	0.002
8	−0.666	0.576	0.331
9	−1.423	−0.181	0.033
10	−0.717	0.525	0.276
Average	−1.242		0.718

So $\alpha = -1.242$, and the variance of the residuals is 0.718.

Finally, asserting the presumption of normality:

$$Y_i = -1.242 + 1.873^*X_{1i} - 1.221^*X_{2i} + \varepsilon_i, \ \varepsilon_i \sim N(0, 0.718)$$

Note that the variance for ε_i is biased and the unbiased figure is this figure multiplied by $\frac{10}{10-3} = 1.429$ or 1.206.

b.

Trial Number	y	ŷ	$(y - \bar{y})^2$	$(\hat{y} - \bar{y})^2$
1	−5.760	−6.089	8.626	10.664
2	0.030	−0.523	8.140	5.290
3	−0.250	−0.873	6.620	3.802
4	−2.720	−0.347	0.011	6.131
5	−3.080	−3.254	0.066	0.185
6	−7.100	−6.834	18.293	16.085
7	−4.100	−4.142	1.631	1.741
8	0.140	−0.436	8.779	5.698
9	−6.130	−5.949	10.936	9.774
10	0.740	0.215	12.695	9.227
Average	−2.823			
		SUM	75.797	68.598

So, the $TSS = 75.797$ and the $ESS = 68.598$.

$$R^2 = \frac{ESS}{TSS} = \frac{68.598}{75.797} = 90.5\%$$

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}(1 - R^2)$$

$$= 1 - \frac{10-1}{10-2-1}(1 - 0.905) = 0.878$$

$$F = \frac{(RSS_R - RSS_U)/q}{RSS_U/(N - k_U - 1)} = \frac{(7.58 - 0.72)/2}{0.72/(10 - 2 - 1)} = 33.47$$

9

Regression Diagnostics

Learning Objectives

After completing this reading you should be able to:

- Explain how to test whether a regression is affected by heteroskedasticity.
- Describe approaches to using heteroskedastic data.
- Characterize multicollinearity and its consequences; distinguish between multicollinearity and perfect collinearity.
- Describe the consequences of excluding a relevant explanatory variable from a model and contrast those with the consequences of including an irrelevant regressor.
- Explain two model selection procedures and how these relate to the bias-variance tradeoff.
- Describe the various methods of visualizing residuals and their relative strengths.
- Describe methods for identifying outliers and their impact.
- Determine the conditions under which OLS is the best linear unbiased estimator.

This chapter examines model specification. Ideally, a model should include all variables that explain the dependent variable and exclude all that do not. In practice, achieving this goal is challenging since the true model generating the data is unknowable, and the model selection process must account for the costs and benefits of including additional variables. Once a model has been selected, the specification should be checked for any obvious deficiencies. Standard specification checks include residual diagnostics and formal statistical tests to examine whether the assumptions used to justify the OLS estimator(s) are satisfied.

Omitting explanatory variables that affect the dependent variable creates biased coefficients on the included variables. The bias depends on both the coefficient of the excluded variable and the correlation between the excluded variable and the remaining variables in the model. When a variable is excluded, the coefficients on the included variables adjust to capture the variation shared between the excluded and the included variables.

On the other hand, including irrelevant variables does not bias coefficients. In large samples, the coefficient on an extraneous variable converges to its population value of zero. Including an unnecessary explanatory variable does, however, increase the uncertainty of the estimated model parameters and their standard errors. When the explanatory variables are correlated, the extraneous variable competes with the relevant variables to explain the data, which in turn makes it more difficult to precisely estimate the parameters that are non-zero.

Determining whether a variable should be included in a model reflects a bias-variance tradeoff. Large models that include all conceivable explanatory variables are likely to have coefficients that are unbiased (i.e., equal, on average, to their population values). However, because it may be difficult to attribute an effect to a single variable, the coefficient estimators in large models are also relatively imprecise, particularly if the sample size is small. Smaller models, on the other hand, lead to more precise estimates of the effects of the included explanatory variables.

Model selection captures this bias-variance tradeoff when sample data are used to determine whether a variable should be included. This chapter introduces two methods that are widely used to select a model from a set of candidate models: general-to-specific model selection and cross-validation.

Once a model has been chosen, additional specification checks are used to determine whether the specification is accurate. Any defects detected in this post-estimation analysis are then used to adjust the model. The standard specification checks include testing to ensure that the functional form is adequate, the model parameters are constant, and the assumption that the errors have constant variance (i.e., homoscedasticity) is compatible with the data. Tests of functional form or parameter stability provide guidance about whether additional variables should be included in the model. Rejecting homoscedasticity requires adjusting the estimator of parameter standard errors, but not the specification of the regression model or the estimated parameters.

The chapter concludes by examining the strengths of the OLS estimator. When five assumptions are satisfied, the OLS estimator is the most accurate in the class of linear unbiased estimators. When model residuals are also normally distributed, then OLS is the best estimator of the model parameters (in the sense of producing the most precise estimates of model parameters among a wide range of candidate estimators).

9.1 OMITTED VARIABLES

An omitted variable is one that has a non-zero coefficient in the true model generating the data but is not included in an estimated model. Omitting a variable has two effects.

1. First, the remaining variables absorb the effects of the omitted variable attributable to common variation. This changes the regression coefficients on the included variables so that they do not consistently estimate the effect of a change in the explanatory variable on the dependent variable (holding all other effects constant).
2. Second, the estimated residuals are larger in magnitude than the true shocks. This is because the residuals contain both the true shock *and* any effect of the omitted variable that cannot be captured by the included variables.

Suppose the data generating process includes variables X_1 and X_2:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

Now suppose X_2 is excluded from the estimated model. The model fit is then:

$$Y_i = \alpha + \beta_1 X_{1i} + \epsilon_i$$

In large samples, the OLS estimator $\hat{\beta}_1$ converges to:

$$\beta_1 + \beta_2 \delta$$

where:

$$\delta = \frac{\text{Cov}[X_1, X_2]}{\text{V}[X_1]}$$

is the population value of the slope coefficient in a regression of X_2 on X_1.

The bias due to the omitted variable X_2 depends on the population coefficient of the excluded variable β_2 and the strength of the relationship between X_1 and X_2 (which is measured by δ). When X_1 and X_2 are highly correlated, then X_1 can explain a substantial portion of the variation in X_2 and the bias is large. If X_1 and X_2 are uncorrelated (so that $\delta = 0$), then $\hat{\beta}_1$ is a consistent estimator of β_1.

In other words, omitted variables bias the coefficient on any variable that is correlated with the omitted variable(s). Note that economic data are generally correlated (often highly), and so omitting relevant variables invariably creates biased and inconsistent estimates of the effect of the included variables.

9.2 EXTRANEOUS INCLUDED VARIABLES

An extraneous (also known as a superfluous or irrelevant) variable is one that is included in the model but is not needed. This type of variable has a true coefficient of 0 and is consistently estimated to be 0 in large samples.

However, there is a cost to including unnecessary variables. To see this, recall the definition of the adjusted R^2 from Chapter 8:

$$\bar{R}^2 = 1 - \xi \frac{RSS}{TSS},$$

where the adjustment factor is: $\xi = (n - 1)/(n - k - 1)$.

Adding variables increases k, which increases ξ and, therefore, reduces $\bar{R}^2$. Meanwhile, an expanded model nearly always has a smaller *RSS*, which (sometimes) offsets the increase in ξ to produce a larger $\bar{R}^2$. However, this is *not* true when the true coefficient is equal to 0. In that case, the *RSS* remains the same even as ξ grows, resulting in a smaller $\bar{R}^2$ as well as a larger standard error.

Furthermore, the standard error grows larger as the correlation between X_1 and X_2 increases. In most applications to financial data, correlations are large and so the impact of an extraneous variable on the standard errors of relevant variables can be substantial.

9.3 THE BIAS-VARIANCE TRADEOFF

The choice between omitting a relevant variable and including an irrelevant variable is ultimately a tradeoff between bias and variance. Larger models tend to have a lower bias (due to the inclusion of additional relevant variables), but they also have less precise estimated parameters (because including extraneous variables increases estimation error). On the other hand, models with few explanatory variables have less estimation error but are more likely to produce biased parameter estimates.

The bias-variance tradeoff is *the* fundamental challenge in variable selection.[1] There are many methods for choosing a final model from a set of candidate explanatory variables, and some of these methods always select the true model (at least in large samples). Other methods perform well in applications of linear regression despite having weaker theoretical foundations.

Two of these methods are general-to-specific model selection and *m*-fold cross-validation.[2] General-to-specific (GtS) model selection begins by specifying a large model that includes all potentially relevant variables based on some preexisting theory or intuition. If there are coefficients in this estimated model that are statistically insignificant (using a test size of α), the variable with the coefficient having the smallest absolute *t*-statistic is deemed empirically irrelevant and is therefore removed. Then, the model is re-estimated using the remaining explanatory variables. These two steps (remove and re-estimate) are repeated until the model contains no coefficients that are statistically insignificant. Common choices for α are between 1% and 0.1% (which correspond to selecting models where all absolute *t*-statistics are at least 2.57 or 3.29, respectively).

m-fold cross-validation is an alternative method that is popular in modern data science. It is designed to select a model that performs well in fitting observations not used to estimate the parameters (a process known as out-of-sample prediction). Cross-validation selects variables with consistent predictive power and excludes variables with small coefficients that do not improve out-of-sample predictions. This mechanism has the effect of directly embedding the bias-variance tradeoff in the model selection, as the model with the lowest forecasting mean squared error will be the one that includes all of the empirically relevant variables and excludes all of the irrelevant variables.

The first step is to determine a set of candidate models. When the number of explanatory variables is small, then it is possible to consider all combinations of explanatory variables. In a data set with ten explanatory variables, for example, $2^{10} = 1{,}024$ distinct candidate models can be constructed.[3]

Cross-validation begins by randomly splitting the data into m equal sized blocks. Common choices for m are five and ten. Parameters are then estimated using $m - 1$ of the m blocks, and the residuals are computed using these estimated parameter values on the data in the excluded block. These residuals are known as out-of-sample residuals, because they are not included in the sample used to estimate parameters. This

[1] Recall that bias measures the expected deviation between an estimator and the true value of the unknown coefficient and is defined $\text{Bias}(\hat{\theta}) = \text{E}[\hat{\theta}] - \theta_0$. Variance measures the squared deviation, $\text{V}[\hat{\theta}] = \text{E}[(\hat{\theta} - \text{E}[\hat{\theta}])^2]$, and so bias and variance are not directly comparable. In practice, the bias-variance tradeoff is a tradeoff between squared bias and variance because $\text{E}[(\hat{\theta} - \theta_0)^2] = \text{Bias}^2(\hat{\theta}) + \text{V}[\hat{\theta}]$.

[2] It most other references on model selection, this type of cross-validation is referred to as *k*-fold cross-validation. *m* is used here to clarify that the number of blocks in the cross-validation (m) is distinct from the number of explanatory variables in the model (k).

[3] If a data set has n candidate explanatory variables, then then there are 2^n possible model specifications. When n is larger than 20, the number of models becomes very large, and so it is not possible to consider all specifications.

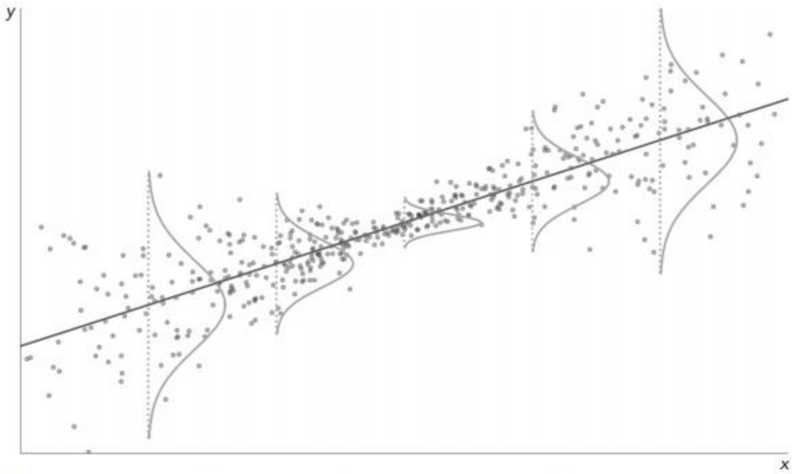


Figure 9.1 An example of heteroskedastic data. The standard deviation of the errors is increasing in the distance between X and $E[X]$, so that the observations at the end of the range are noisier than those in the center.

process of estimating parameters and computing residuals is repeated a total of *m* times (so that each block is used to compute residuals once).

The sum of squared errors is then computed using the residuals estimated from the out-of-sample data. Finally, the preferred model is chosen from the set of candidate models by selecting the model with the smallest out-of-sample sum of squared residuals. In machine learning or data science applications, the $m - 1$ blocks are referred to as the *training set* and the omitted block is called the *validation set*.

9.4 HETEROSKEDASTICITY

Homoscedasticity is one of the five assumptions used to determine the asymptotic distribution of an OLS estimator. It requires that the variance of ϵ_i is constant and so does not systematically vary with any of the explanatory variables. When this is not the case, then the residuals are *heteroskedastic*.

Formally, homoscedasticity requires that the conditional variance of the shock is constant (i.e., $V[\epsilon_i | X_{1i}, X_{2i}, \ldots, X_{ki}] = \sigma^2$). In models of financial data, this assumption is often false. A consequence of heteroskedasticity is that the asymptotic distribution of the estimated parameters takes a different form. However, heteroskedasticity *does not* affect the consistency or unbiasedness of the OLS parameter estimator.

Figure 9.1 contains a simulated data set with heteroskedastic shocks. The residual variance increases as X moves away from its mean; consequently, the largest variances occur when $(X - \mu_X)^2$ is largest. This form of heteroskedasticity is common in financial asset returns.

In a regression, the most extreme observations (of X) contain the most information about the slope of the regression line. When these observations are coupled with high-variance residuals, estimating the slope becomes more difficult. This is reflected by a larger standard error for $\hat{\beta}$.

When residuals are heteroskedastic, the standard errors can be estimated using White's estimator (also called Eiker-White in some software packages). This incorporates a modification to the usual formula for standard errors that considers the form of the heteroskedasticity, and so they are known as heterosketasticty-robust standard errors. Parameters can then be tested using *t*-tests by using White's standard error in the place of standard error used for homoscedastic data. On the other hand, *F*-tests, are not as easy to adjust for heteroskedasticity and so caution is required when testing multiple hypotheses if the shocks are heteroskedastic.[4]

[4] A Wald test can be used when residuals are heteroskedastic. The expression for this type of test requires linear algebra. This modification of *F*-tests is automatically employed by statistical software when using heteroskedasticity-robust parameter covariances (i.e., when using White standard errors).

A simple test for heteroskedasticity, proposed by White is implemented as a two-step procedure.[5]

1. Estimate the model and compute the residuals $(\hat{\epsilon}_i)$.
2. Regress the *squared* residuals on a constant, all explanatory variables, and the cross-product of all explanatory variables (including the product of each variable with itself).

For example, if the original model has two explanatory variables:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

then the first step computes the residuals using the OLS parameter estimators:

$$\hat{\epsilon}_i = Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}$$

The second step regresses the squared residuals on a constant, the original explanatory variables, their squares, and their cross-products. In this example, the variables would be: X_1, X_2, X_1^2, X_2^2, and X_1X_2:

$$\hat{\epsilon}_i^2 = \gamma_0 + \gamma_1 X_{i1} + \gamma_2 X_{2i} + \gamma_3 X_{1i}^2 + \gamma_4 X_{2i}^2 + \gamma_5 X_{1i} X_{2i} + \eta_i \quad (9.1)$$

The second stage is sometimes known as an *auxiliary regression.*

If the data are homoscedastic, then $\hat{\epsilon}_i^2$ should not be explained by any the variables on the right-hand side and the null is therefore

$$H_0: \gamma_1 = \cdots = \gamma_5 = 0$$

The test statistic is computed as nR^2, where the R^2 is obtained from the second regression. The test statistic has a $\chi^2_{k(k+3)/2}$ distribution, where k is the number of explanatory variables in the first-stage model.[6]

When $k = 2$ as in the example above, the null imposes five restrictions and the test statistic is distributed χ^2_5. Large values of the test statistic indicate that the null is false, so that homoscedasticity is rejected in favor of heteroskedasticity. When this occurs, heteroskedasticity-robust (White) standard errors should be used.

Table 9.1 contains the results of White's test on the industry portfolios considered in Chapter 8. The first-step model is derived from CAPM:

$$R_{p^*i} = \alpha + \beta R_{m^*i} + \epsilon_i,$$

where R_{p^*i} is the excess return on a portfolio and R_{m^*i} is the excess return on the market.

[5] White, H, "A Heteroskedasticity-consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica*, 48(4) (1980): 817–838.

[6] The Breusch-Pagan test is a special case of White's test that excludes the cross-products of the explanatory variables, and so the distribution of the test statistic is a χ^2_{2k}. This version of the test is often preferred in large models to moderate the number of regressors in the second-step regression.

The residuals from the regression are constructed as:

$$\hat{\epsilon}_i = R_{p^*i} - \hat{\alpha} - \hat{\beta} R_{m^*i}$$

These residuals are then squared and regressed on a constant, the excess market return, and the squared excess market return:[7]

$$\hat{\epsilon}_i^2 = \gamma_0 + \gamma_1 R_{m^*i} + \gamma_2 (R_{m^*i})^2 + \eta_i$$

Table 9.1 reports the parameter estimates for γ_1 and γ_2 (which should be 0 if the data are homoscedastic). The values below the estimated coefficients are *t*-statistics.

Apart from the banking industry, all estimates of γ_1 are statistically insignificant at the 10% level (i.e., they are smaller in magnitude than 1.645). This indicates that the market return does not generally linearly affect the variance of the residuals.

Table 9.1 White's Test for Heteroskedasticity Applied to the Industry Portfolios. The First-Step Model Is a CAPM. The Second-Step Model Is $\hat{\epsilon}_i^2 = \gamma_0 + \gamma_1 R_{m^*i} + \gamma_2 (R_{m^*i})^2 + \eta_i$, where R_{m^*i} Is the Excess Return on the Market. The Values Below the Estimated Coefficients Are *t*-Statistics. The Third Column Contains White's Test Statistic, nR^2 from the Second Stage Regression, Which Has a χ^2_2 Distribution. The p-Value of the Test Statistic Is in Parentheses in This Column. The Final Column Reports the R^2 from This Regression. The Number of Observations is $n = 360$

	γ_1	γ_2	White Stat.	R^2
Banking	−0.488	0.140	40.63	0.113
	(−2.67)	(5.50)	(0.000)	
Shipping Containers	0.144	0.053	0.95	0.003
	(0.36)	(0.96)	(0.623)	
Chemicals	0.042	0.096	7.72	0.021
	(0.17)	(2.77)	(0.021)	
Electrical Equipment	0.191	0.092	10.21	0.028
	(0.93)	(3.23)	(0.006)	
Computers	0.348	0.205	10.70	0.030
	(0.78)	(3.32)	(0.005)	
Consumer Goods	0.492	0.085	3.51	0.010
	(1.29)	(1.61)	(0.173)	
Retail	0.102	0.050	3.10	0.009
	(0.50)	(1.76)	(0.212)	

[7] Note there are no cross-product terms in this case since there is only a single explanatory variable in the first stage regression.

Table 9.2 **Parameter Estimates from Regressions of the Returns of Each Industry Portfolio on the Returns of the Factor Portfolios: The Market, the Size Factor, and the Value Factor. The Line Immediately Below the Parameter Estimates Contains the *t*-Statistics of the Coefficient Computed Using White's Heteroskedasticity-Robust Covariance Estimator. The Next Line Contains the *t*-Statistics (in italics) of the Coefficient That Are Only Valid if the Shocks Are Homoscedastic. The Final Two Columns Report Test Statistics of the Null $H_0: \beta_s = \beta_v = 0$ Using the Standard *F*-Test and the Heteroskedasticity-Robust *F*-Test. The Number Below Each Test Statistic Is the p-Value of the Test Statistic**[8]

	α	β_m	β_s	β_v	F-test	Robust F
Banking	−0.172	1.228	−0.156	0.838	135.3	87.2
	(−1.019)	(23.015)	(−2.729)	(12.124)	(0.000)	(0.000)
	(−1.095)	*(32.261)*	*(−3.027)*	*(15.152)*		
Shipping Containers	−0.003	1.011	−0.012	0.276	6.57	4.11
	(−0.012)	(17.128)	(−0.107)	(2.851)	(0.002)	(0.016)
	(−0.012)	*(18.633)*	*(−0.161)*	*(3.510)*		
Chemicals	−0.053	1.098	−0.028	0.458	32.1	18.0
	(−0.309)	(21.988)	(−0.433)	(5.783)	(0.000)	(0.000)
	(−0.314)	*(26.902)*	*(−0.506)*	*(7.720)*		
Electrical Equipment	0.103	1.263	−0.033	0.115	2.21	1.53
	(0.585)	(25.147)	(−0.453)	(1.653)	(0.111)	(0.217)
	(0.585)	*(29.613)*	*(−0.570)*	*(1.855)*		
Computers	0.055	1.296	0.224	−0.666	40.8	24.7
	(0.228)	(18.343)	(2.285)	(−6.369)	(0.000)	(0.000)
	(0.226)	*(21.988)*	*(2.800)*	*(−7.784)*		
Consumer Goods	0.215	0.675	−0.174	0.117	9.18	1.61
	(1.372)	(11.779)	(−1.306)	(1.507)	(0.000)	(0.200)
	(1.311)	*(17.026)*	*(−3.241)*	*(2.028)*		
Retail	0.221	0.927	−0.036	0.010	0.27	0.12
	(1.346)	(20.006)	(−0.478)	(0.154)	(0.763)	(0.887)
	(1.347)	*(23.327)*	*(−0.661)*	*(0.174)*		

However, most of the coefficients on γ_2 are statistically significant at the 10% level, and so the shock variance is large when the squared market return is large; in other words, the volatilities of the shock and of the market are positively related.

The third column reports the test statistic for White's test (which is nR^2 from the second-stage regression on the residuals). The null of homoscedasticity imposes two restrictions on the coefficients in the second-stage regression ($H_0: \gamma_1 = \gamma_2 = 0$), and so the test statistic has a χ^2_2 distribution. Four of the seven series reject the null hypothesis of homoskedasticity and therefore appear to be heteroskedastic. The final column reports the R^2 of the second-stage model. All regressions use 30 years of monthly observations ($n = 360$).

Table 9.2 includes *t*-statistics computed using White's covariance estimator in the three-factor models. It contains the estimated parameters for each industry portfolio, along with the heteroskedasticity-robust *t*-statistics and the *t*-statistics computed assuming homoscedasticity (in *italics*).

The *t*-statistics on the size and value factor coefficients that account for heteroskedasticity are generally smaller than the *t*-statistics computed assuming homoscedasticity. The changes in the consumer goods industry portfolio are large enough to alter the conclusion about the significance of the size factor (because the *t*-statistic decreases from −3.2 to −1.3).

[8] The standard definition of the heteroskedasticity-robust *F*-statistic has a χ^2_q distribution, while the standard *F*-test has a $F_{q,n-k-1}$ distribution. When *n* is large, $F_{q,n-k-1} \approx \chi^2_q/q$, and so the robust test statistic has been divided by $q = 2$ to make the values of these two test statistics approximately comparable.

The final two columns contain test statistics for the null that CAPM is adequate (i.e., $H_0: \beta_s = \beta_v = 0$). The two test statistics are the *RSS*-based *F*-test statistic (which is only valid when the data are homoscedastic) and a heteroskedasticity-robust *F*-test statistic. Below each test statistic is its p-value.

The robust *F*-tests show a similar effect: The robust test statistics are smaller and have larger p-values than the test statistics assuming that the shocks are homoscedastic. In the test of the coefficients in the consumer goods portfolio, this difference changes the conclusion: Whereas, the standard *F*-test indicates that the null is false, the robust version has a p-value of 20% and so the null cannot be rejected. This test result is consistent with the change in the *t*-statistic of the coefficients for this industry. This is a common pattern when relaxing the assumption of homoscedasticity; heteroskedasticity-robust standard errors indicate that parameters are less precisely estimated in the presence of heteroskedasticity and therefore the standard errors need to be increased to reflect this, resulting in smaller (in absolute value) t-ratios.

Approaches to Modeling Heteroskedastic Data

There are three common approaches used to model data with heteroskedastic shocks. The first (and simplest) is to ignore the heteroskedasticity when estimating the parameters and then use the heteroskedasticity-robust (White) covariance estimator in hypothesis tests. This approach is frequently used because the parameters are estimated using OLS. However, while this method is simple, it often produces substantially less precise model parameter estimates when compared to methods that directly address the heteroskedasticity.

The second approach is to transform the data. For example, it is common to find heteroskedasticity in data that are strictly positive (e.g., trading volume or firm revenues). When data are always positive, it is tempting to model the natural log of the dependent variable, because it can reduce or even eliminate heteroskedasticity and may provide a better description of the data. An alternative is to transform the data by dividing the dependent variable by another (strictly positive) variable. For example, dividing the dividend per share by the share price produces the yield of a stock.[9]

The final (and most complicated) approach is to use weighted least squares (WLS). WLS is a generalization of OLS that applies a set of weights to the data before estimating parameters. If $V[\epsilon_i] = w_i^2\sigma^2$ and w_i is a variable known to cause the heteroskedasticity, then transforming the data by dividing by w_i removes the heteroskedasticity from the errors. WLS regresses Y_i/w_i on X_i/w_i so that:

$$\frac{Y_i}{w_i} = \alpha\frac{1}{w_i} + \beta\frac{X_i}{w_i} + \frac{\epsilon_i}{w_i},$$
$$\bar{Y}_i = \alpha\bar{C}_i + \beta\bar{X}_i + \bar{\epsilon}_i. \tag{9.2}$$

The parameters of this model are estimated using OLS on the transformed data.[10] The modified error is ϵ_i/w_i, and so the variance of the error ($V[\epsilon_i/w_i] = \sigma^2$) is constant across observations. The challenge with implementing WLS is determining the weights. One approach is to use a three-step procedure, which estimates the model with OLS, then estimates the weights using the residuals, and then finally re-estimates the transformed model. This approach is known as Feasible WLS or Feasible Generalized Least Squares (FWLS/FGLS).

9.5 MULTICOLLINEARITY

In any practical context, the explanatory variables in a regression model will never be completely independent and thus will be correlated to some extent. A problem known as multicollinearity arises if this correlation is too high. More specifically, multicollinearity occurs when one or more explanatory variables can be substantially explained by the other(s). For example, if a model has two explanatory variables, then the variables are multicollinear if the R^2 in the regression of one on the other is very high.

Multicollinearity differs from perfect collinearity, where one of the variables is perfectly explained by the others (in which case the R^2 of a regression of X_j on the remaining explanatory variables is 1). The rule-of-thumb is that R^2 above 90% creates serious issues in moderate sample sizes (i.e., 100s of observations).[11]

Multicollinearity is a common problem in finance and risk management because many regressors are simultaneously determined by and sensitive to the same news. Multicollinearity is not ruled out by any of the assumptions of linear regression, and so it does not pose a technical challenge to parameter estimation or hypothesis testing. It does, however, create challenges in modeling data.

[9] Note that yields are often homoscedastic even when dividends per share are not.

[10] The WLS regression regresses the weighted version of Y_i, namely $\bar{Y}_i$, on two explanatory variables. The first, $\bar{C}_i$, is the inverse of the weight, $1/w_i$. The second is the weighted version of X_i, namely $\bar{X}_i$. While the WLS regression does not explicitly include an intercept, α is still interpretable as the intercept.

[11] In a linear regression with a single explanatory variable, the R^2 is the squared correlation (ρ^2) between the dependent and the explanatory variable. An $R^2 = 90\%$ corresponds to a correlation of 95%.

When data are multicollinear, it is common to find coefficients that are jointly statistically significant (as evidenced by the F-statistic of the regression) and yet have small individual t-statistics (e.g., less than 1.96 in magnitude). This contradiction occurs because the joint statistical analysis can identify some effect from the regressors as a group (using the F-statistic) but cannot uniquely attribute the effect to a single variable (using the t-statistics).

There are two options available to address multicollinear explanatory variables. The first (and easiest) is to ignore the issue because it is technically not a problem. The alternative is to identify multicollinear variables and to consider removing these from the model. The standard method to determine whether variables are excessively multicollinear is to use the variance inflation factor. This measure compares the variance of the regression coefficient on an explanatory variable X_j in two models: one that only includes X_j and one that includes all k explanatory variables:

$$X_{ji} = \gamma_0 + \gamma_1 X_{1i} + \ldots + \gamma_{j-1} X_{j-1i} + \gamma_{j+1} X_{j+1i} + \ldots + \gamma_k X_{ki} + \eta_i \quad (9.3)$$

The variance inflation factor for variable j is then:

$$VIF_j = \frac{1}{1 - R_j^{2'}} \quad (9.4)$$

where R_j^2 comes from a regression of X_j on the other variables in the model. Values above 10 (i.e., which indicate that 90% of the variation in X_j can be explained by the other variables in the model) are considered excessive.

Variables with exceedingly high variance inflation factors should be considered for exclusion from the model. In practice, model selection procedures screen out models with high VIF variables because these variables have parameters that are poorly estimated and excluding them tends to have little impact on the fitted values from the model.

Residual Plots

Residual plots are standard methods used to detect deficiencies in a model specification. An ideal model would have residuals that are not systematically related to any of the included explanatory variables. The residuals should also be relatively small in magnitude (i.e., typically within $\pm 4s$, where s^2 is the estimated variance of the shocks in the model).

The basic residual plot compares $\hat{\epsilon}_i$ (y-axis) against the realization of the explanatory variables x_i. An alternative uses the standardized residuals $\hat{\epsilon}/s$ so that the magnitude of the deviation is more apparent. Both outliers and model specification problems (e.g., omitted nonlinear relationships) can be detected in these plots.

Outliers

Outliers are values that, if removed from the sample, produce large changes in the estimated coefficients. Cook's distance measures the sensitivity of the fitted values in a regression to dropping a single observation j. It is defined as:

$$D_j = \frac{\sum_{i=1}^{n} (\hat{Y}_i^{(-j)} - \hat{Y}_i)^2}{ks^2} \quad (9.5)$$

where $\hat{Y}_i^{(-j)}$ is the fitted value of Y_i when observation j is dropped and the model is estimated using $n - 1$ observations, k is the number of coefficients in the regression model, and s^2 is the estimate of the error variance from the model that uses all observations.

D_j should be small when an observation is an inlier (i.e., does not disproportionately affect parameter estimates). Values of D_j larger than 1 indicate that observation j has a large impact on the estimated model parameters and so is an outlier.

For example, consider the data in Table 9.3, which show an illustrative example with a potential outlier.

Table 9.3 **Sample Data on an Explanatory and an Explained Variable with a Potential Outlier**

Trial	x	y
1	3.86	1.89
2	1.89	0.64
3	1.38	−0.62
4	0.20	1.20
5	−0.81	−2.42
6	1.92	0.67
7	2.96	0.81
8	1.60	0.26
9	1.46	0.72
10	1.52	−0.41
11	0.68	1.13
12	0.03	−1.78
13	1.56	1.46
14	−0.12	−0.62
15	7.60	1.85

Note that Trial 15 yields a result that is quite different from the rest of the data. To see whether this observation is an outlier, a model is fit on the entire data set (i.e., $\hat{Y}_i$) as well as on just the first 14 observations (i.e., $\hat{Y}_i^{(-j)}$). The resulting coefficients are

$$\hat{Y}_i = 1.38 + 1.04X_i$$

and

$$\hat{Y}_i^{(-j)} = 1.15 + 0.69X_i$$

The fitted values from these models are shown in the table below:

Table 9.4 Fitted Values for a Regression Model That Includes All Observations and for a Model Where Observation 15, the Potential Outlier, Is Excluded

Trial	X_i	$\hat{Y}_i$	$\hat{Y}_i^{(-j)}$	$(\hat{Y}_i^{(-j)} - \hat{Y}_j)^2$
1	1.89	3.35	2.45	0.81
2	0.64	2.05	1.59	0.21
3	−0.62	0.74	0.73	0.00
4	1.20	2.63	1.98	0.43
5	−2.42	−1.14	−0.51	0.39
6	0.67	2.08	1.61	0.22
7	0.81	2.23	1.71	0.27
8	0.26	1.65	1.33	0.11
9	0.72	2.13	1.65	0.24
10	−0.41	0.96	0.87	0.01
11	1.13	2.56	1.93	0.40
12	−1.78	−0.47	−0.07	0.16
13	1.46	2.90	2.15	0.56
14	−0.62	0.74	0.73	0.00
15	1.85	3.31	2.42	0.79
			Sum	4.60

The variance from the full run (i.e., s^2) is calculated as 2.19. Calculating Cook's distance yields

$$D_j = \frac{\sum_{i=1}^{n}(\hat{Y}_i^{(-j)} - \hat{Y}_i)^2}{ks^2} = \frac{4.6}{2 \times 2.19} = 1.05$$

Because $D_j > 1$, Trial 15 can be considered an outlier.

As another example, consider the industry portfolio data from Chapter 8. The top panel in Figure 9.2 shows Cook's distance in the three-factor model, where the dependent variable is the excess return on the consumer good's industry portfolio. The plot shows that one observation has a distance larger than 1. The bottom panel shows the effect of this single observation on the estimate of the coefficient on the size factor $\hat{\beta}_s$.

Dropping this single influential observation changes $\hat{\beta}_s$ from −0.30 to −0.18. The standard error of $\hat{\beta}_s$ is 0.054, and so this single observation produces a change of 2.2 standard errors. The other two model parameters do not meaningfully change, because the identified observation was not associated with a large deviation in the other explanatory variables.

9.6 STRENGTHS OF OLS

Under assumptions introduced in Chapter 7, OLS is the Best Linear Unbiased Estimator (BLUE). Best indicates that the OLS estimator achieves the smallest variance among *any* estimator that is linear and unbiased.

OLS IS BLUE

OLS is a linear estimator because, using the bivariate regression setup for illustration, both $\hat{\alpha}$ and $\hat{\beta}$ can be written as linear functions of Y_i, so that:

$$\hat{\beta} = \sum_{i=1}^{n} W_i Y_i$$

The weights, W_i, in a linear estimator cannot depend on the Y_i, but they may depend on other variables. Note that the weights in the OLS estimator of β:

$$W_i = \frac{X_i - \bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

depend only on the values of X.

Meanwhile, the weights in the OLS estimator of α are

$$W_i = \frac{1}{n} - \bar{X}\left(\frac{X_i - \bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right),$$

and so also do not depend on Y.

OLS is the best estimator in the sense that any other LUE must have larger variances. The intuition behind this conclusion follows from a result that any other LUE $\bar{\beta}$ can be expressed as $\hat{\beta} + \hat{\delta}$, where $\hat{\beta}$ is the OLS estimator and $\hat{\delta}$ is uncorrelated with $\hat{\beta}$. The variance of any other LUE estimator is then:

$$V[\bar{\beta}] = V[\hat{\beta}] + V[\hat{\delta}]$$

This expression is minimized when $V[\hat{\delta}] = 0$. Because this only happens when $\bar{\beta}$ is the OLS estimator, any other estimator would have a larger variance. This result is known as the Gauss-Markov Theorem.

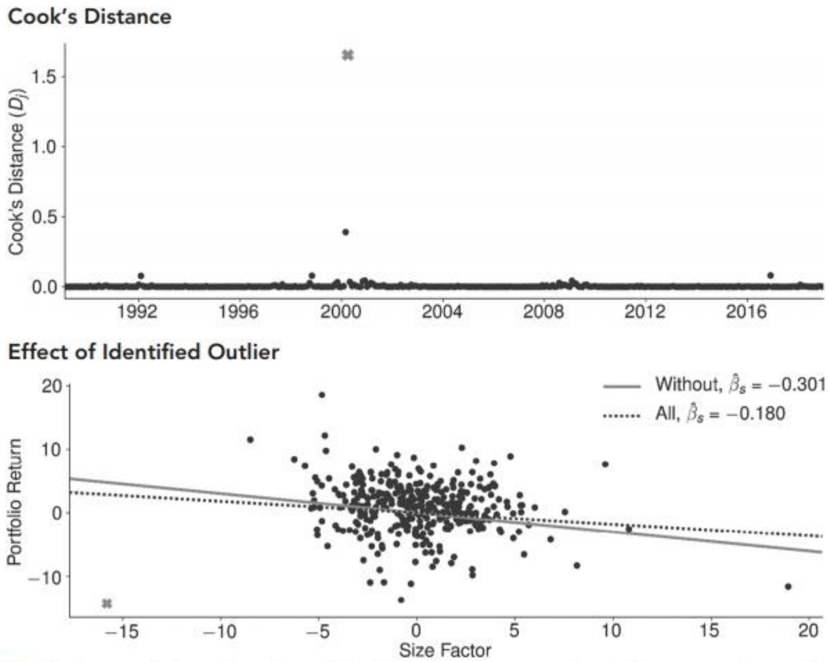


Figure 9.2 The top panel shows the values of Cook's distance in a regression of the excess return on the consumer good industry portfolio on the three factors, the excess return on the market and the returns on the size and value portfolios. The bottom panel shows the effect on the estimated coefficient on the size factor when the outlier identified by Cook's distance is removed.

While BLUE is a strong argument in favor of using OLS to estimate model parameters, it comes with two important caveats.

1. First, the BLUE property of OLS depends crucially on the homoscedasticity of the residuals. When the residuals have variances that change with X (as is common in financial data), then it is possible to construct better LUEs of α and β using WLS (under some additional assumptions).
2. Second, many estimators are not linear. For example, many maximum likelihood estimators (MLEs) are nonlinear.[12] MLEs have desirable properties and generally have the smallest asymptotic variance of any consistent estimator. MLEs are, however, generally biased in finite samples and so it is not possible to compare them to OLS in fixed sample sizes.

If $\epsilon_i \overset{iid}{\sim} N(0,\sigma^2)$, then OLS is the Best Unbiased Estimator (BUE). That is, the assumption that the errors are normally distributed strengthens BLUE to BUE so that OLS is the Best (in the sense of having the smallest variance) Unbiased Estimator among all linear *and* nonlinear estimators. This is a compelling justification for using OLS when the errors are iid normal random variables.[13]

However, it is not the case that OLS is only applicable when the errors are normally distributed. The only requirements on the errors are the assumptions that the residuals have conditional mean zero and that there are no outliers. Normality of the errors

[12] Maximum likelihood estimators are another class of estimators.

[13] The standard test of normality is the Jarque-Bera statistic, which tests two properties of normal random variables: that the skewness is zero and the kurtosis is 3.

is *not* required for $\hat{\alpha}$ and $\hat{\beta}$ to accurately estimate the population values or for the estimators to have desirable properties (e.g., consistency or asymptotic normality).

9.7 SUMMARY

Developing a model requires making choices that reflect a bias-variance tradeoff. Large models are less likely to omit important explanatory variables, and thus, parameter estimators are less biased. However, large models are also more likely to include irrelevant variables that reduce the precision of the coefficient estimators. Large models are also more likely to contain variables that are redundant (i.e., whose effects can be explained by other variables in the model). General-to-specific model selection and cross-validation are two methods that account for the bias-variance tradeoff to select a single specification.

An accurate model should have several desirable features. First, included variables should not be excessively collinear. The variance inflation factor is a simple method to check for excess collinearity and provides guidance about which variables can be removed.

Second, the impact of outliers should be recognized. Graphical tools provide a simple method to inspect residuals for outliers or misspecification. Cook's distance provides a numerical measure of the outlyingness of a data point. If outliers are detected, the validity of the data point should be considered. In applications in finance and risk management, especially when modeling return data, dropping these points could be misleading and result in an underestimation of risk because the return occurred in the past and similar returns may occur in the future.

The chapter concludes by describing the conditions where OLS is a desirable method to estimate model parameters. When five key assumptions are satisfied, the OLS parameter estimator is the Best Linear Unbiased Estimator (BLUE), which provides a strong justification for choosing OLS over alternative methods. When the residuals are also normally distributed, the BLUE property is strengthened so that OLS is the BUE (best-unbiased estimator) among all linear and nonlinear estimators.

The following questions are intended to help candidates understand the material. They are *not* actual FRM exam questions.

QUESTIONS

Short Concept Questions

9.1 Define homoscedasticity and heteroskedasticity. When might you expect data to be homoscedastic?

9.2 What conditions are required for a regression to suffer from omitted variable bias?

9.3 What are the costs and benefits of dropping highly collinear variables?

9.4 If you use a general-to-specific model selection with a test size of α, what is the probability that one or more irrelevant regressors are included when the regressors are uncorrelated?

9.5 Why does a variable with a large variance inflation factor indicate that a model may be improved?

9.6 The sample mean is an OLS estimator of the model $Y_i = \alpha + \epsilon_i$. What does the BLUE property imply about the mean estimator?

9.7 What is the strongest justification for using OLS to estimate model parameters?

Practice Questions

9.8 In a model with a single explanatory variable, what value of the R^2 in the second step in a test for heteroskedasticity indicates that the null would be rejected for sample sizes of 100, 500, or 2500? (Hint: Look up the critical values of a χ^2_q, where q is the number of restrictions in the test.)

9.9 Suppose that the true relationship between Y and two explanatory variables is $Y_i = 2 + 1.2X_{1i} - 2.1X_{2i} + \epsilon_i$.

a. What is the population value of β_1 in the regression $Y_i = \alpha + \beta_1 X_1 + \epsilon_i$ if $\rho_{X_1 X_2} = 0.6$, $\sigma^2_{X_1} = 1$ and $\sigma^2_{X_2} = 1/2$?

b. What value of $\rho_{X_1 X_2}$ would make $\beta_1 = 0$?

9.10 In a model with two explanatory variables, $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$, what does the correlation need to be between the two regressors, X_1 and X_2, for the variance inflation factor to be above 10?

9.11 Given the data set:

	Y	X_1	X_2
1	−2.353	−0.409	−0.008
2	−0.114	0.397	−1.216
3	−1.665	−0.856	−0.911
4	−0.364	1.682	0.366
5	−0.081	0.455	−0.639
6	−0.735	−1.39	−1.086
7	−2.507	0.954	0.67
8	−1.144	1.021	0.238
9	−2.419	−0.156	−0.055
10	−3.151	1.382	1.148
11	−2.085	−0.562	−0.135
12	−2.972	−1.554	−0.299
13	−0.633	−1.123	−1.027
14	−2.678	−0.124	0.331
15	−7.095	0.284	2.622

a. Find the parameters for the model:

$$Y = \alpha + \beta_1 X_1.$$

b. Find the parameters for the model:

$$Y = \alpha + \beta_2 X_2.$$

c. Using the results from parts a and b, and the correlation between the two explanatory variables, estimate the parameters for the full model:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2.$$

d. Check the answer to part c by using the Excel function LINEST or the Excel regression macro.

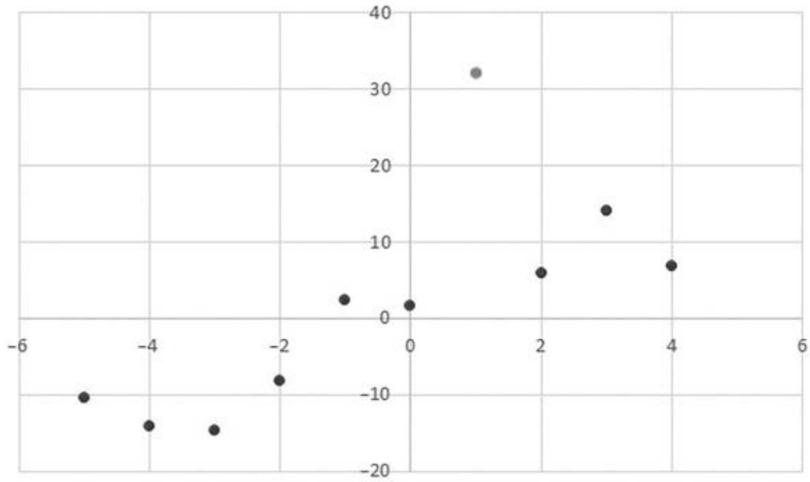
9.12 A given set of data was divided into three equal parts. Three separate models were developed—each model using two of the three parts for fitting. Errors were calculated for each model. The diagram below shows the standard errors for each model run (the orange highlights where the data was used for fitting the model versus the blue indicating the data that was held back):

	M_1	M_2	M_3
1	2.1	0.6	1.1
2	−0.1	−0.4	2.5
3	−0.8	−0.1	1
4	−2.6	−0.5	−0.1
5	−1.1	−0.6	1.3
6	−0.4	0.9	0.5
7	0.6	0	0
8	0.9	−1	−0.4
9	1.2	0.8	−0.1

Using the principles of *m*-fold cross validation, which model should be selected?

9.13 Consider the following data provided in the table and plotted below:

	Y	X
1	−10.42	−5
2	−14.12	−4
3	−14.72	−3
4	−8.25	−2
5	2.3	−1
6	1.59	0
7	5.87	2
8	14.13	3
9	6.78	4
10	32	1



Management is concerned that the last data point in the table is an outlier.

a. Calculate Cook's Distance for the last data point. Use built-in regression functions in Excel to assist as needed, such as LINEST.

b. What is the interpretation of this result? What does it say about the status of the last data point as an outlier?

ANSWERS

Short Concept Questions

9.1 Homoscedasticity is a property of the model errors where they have constant variance. Heteroskedasticity is a property of the errors where their variance changes systematically with the explanatory variables in the model. Experimental data are highly likely to be homoscedastic. In general, the more homogeneous the data, the more likely the errors will be homoscedastic. In finance, we often use data with substantially different scales, for example, corporate earnings or leverage ratios. This heterogeneity is frequently accompanied by heteroskedasticity in model errors.

9.2 There are two conditions. First, the variable must be omitted in the sense that it has a non-zero coefficient in the correct model. Second, the omitted variable must be correlated with an included variable. If the X variables are uncorrelated, then the coefficients can be estimated consistently even if there is an omitted variable.

9.3 The cost is that the model may be misspecified (omitted variable bias) if the dropped variable should be in the true model. The benefit is that the parameter estimates of the remaining variables will be more precise because the OLS estimator will be able to clearly attribute movement in the left-hand-side variable to the remaining variables.

9.4 If the regressors are uncorrelated, then their joint distribution has no correlation, and so the probability that one-or-more irrelevant regressors is included is 1 minus the probability that none are. The probability of including a single irrelevant regressor is $1 - \alpha$, where α is the test size. This is the definition of the size of a test. The probability that no irrelevant regressors are included is $(1 - \alpha)^p$, where p is the number of variables considered.

9.5 When a variable has a large VIF, it is highly correlated with the other variables in the model. The parameter on this variable is usually poorly estimated, and so dropping it tends to have little impact on the overall fit of the model (although it may have a considerable effect on the estimates of the remaining individual parameters).

9.6 The sample mean estimator is the BLUE because it is a special case of OLS if the data are homoscedastic.

9.7 When the errors are iid normally distributed, then the OLS estimator of the regression parameters (α and β, but not s^2) is an MVUE (Minimum Variance Unbiased Estimator), and so there is not a better estimator available.

Solved Problems

9.8 When there is a single explanatory variable in the original model, the auxiliary regression used for White's test will have an intercept and two explanatory variables—the original variable and the variable squared. White's test statistic is nR^2, and the null has a χ^2_2 distribution. When using a test with a 5% size, the critical value is 5.99. Solving for the R^2, $R^2 \leq \frac{5.99}{n}$. The maximum value of R^2 that would not reject is 0.0599, 0.011, and 0.0023 when n is 100, 500, and 2500, respectively.

9.9 **a.** Using the omitted variable formula, when X_2 is omitted from the regression,

$$\beta_1 = 1.2 - 2.1 \times \left(\frac{0.6 \times \sqrt{1} \times \sqrt{1.2}}{1}\right) = -0.18,$$

where $\frac{0.6 \times \sqrt{1} \times \sqrt{1.2}}{1}$ is the regression slope from a regression of X_2 on X_1.

b. The omitted variable formula shows that when a variable is excluded, the coefficient on the included variables is $\beta_1 + \gamma\beta_2$, where γ is the population regression coefficient from the regression $X_{2i} = \delta + \gamma X_{1i} + \eta_i$. This coefficient depends on the correlation between the two and the standard deviations of the two variables, so that $\gamma = \rho_{X_1X_2}\frac{\sigma_{X_2}}{\sigma_{X_1}} = 0.6\frac{\sqrt{1/2}}{\sqrt{1}} = 0.424$. The coefficient is then $1.2 - 2.1 \times 0.424 = 0.309$.

Solving $1.2 - 2.1 \times \left(\frac{\rho \times \sqrt{1} \times \sqrt{1.2}}{1}\right) = 0$ so that if $\rho = \frac{1.2}{2.1 \times \sqrt{1.2}} = 0.522$ then β_1 would be 0 in the omitted variable regression.

9.10 The variance inflation factor is $\frac{1}{1 - R_j^2}$, where R_j^2 measures how well variable j is explained by the other variables in the model. Here there are two variables, and so $R_j^2 = \rho^2_{X_1X_2}$. The variance inflation factor of 10 solves to: $10 = 1/(1 - \rho^2_{X_1X_2})$, so that $(1 - \rho^2_{X_1X_2}) = 1/10$ or $\rho^2_{X_1X_2} = 1 - 1/10 = 0.9$. Correlations greater than (in absolute value) $\sqrt{0.9} \approx 0.948$ would produce values of the variance inflation factor above 10.

9.11 **a.** Using the standard formula:

$$\hat{\beta}_1 = \frac{Cov(Y, X_1)}{V(X_1)}$$

	Y	X1	$(X_1 - \overline{X_1})$	$(Y - \overline{Y})$	$(X_1 - \overline{X_1})(Y - \overline{Y})$	$(X_1 - \overline{X_1})^2$
1	−2.353	−0.409	−0.409	−0.353	0.145	0.167
2	−0.114	0.397	0.397	1.886	0.749	0.158
3	−1.665	−0.856	−0.856	0.335	−0.287	0.733
4	−0.364	1.682	1.682	1.636	2.751	2.829
5	−0.081	0.455	0.455	1.919	0.873	0.207
6	−0.735	−1.39	−1.390	1.265	−1.758	1.932
7	−2.507	0.954	0.954	−0.507	−0.484	0.910
8	−1.144	1.021	1.021	0.856	0.874	1.042
9	−2.419	−0.156	−0.156	−0.419	0.065	0.024
10	−3.151	1.382	1.382	−1.151	−1.591	1.910
11	−2.085	−0.562	−0.562	−0.085	0.048	0.316
12	−2.972	−1.554	−1.554	−0.972	1.511	2.415
13	−0.633	−1.123	−1.123	1.367	−1.535	1.261
14	−2.678	−0.124	−0.124	−0.678	0.084	0.015
15	−7.095	0.284	0.284	−5.095	−1.447	0.081
Average	**−2.000**	**0.000**			**0.000**	**0.933**

Therefore:

$$\hat{\beta}_1 = \frac{0.000}{0.933} = 0.000$$

And, of course:

$$\hat{\alpha} = \overline{Y} - \hat{\beta}_1\overline{X_1} = -2.000$$

So:

$$Y = -2.000$$

b. Proceeding as in part a gives:

	Y	X2	$(X_2-\overline{X_2})$	$(Y-\overline{Y})$	$(X_2-\overline{X_2})(Y-\overline{Y})$	$(X_2-\overline{X_2})^2$
1	−2.353	−0.008	−0.008	−0.353	0.003	0.000
2	−0.114	−1.216	−1.216	1.886	−2.293	1.478
3	−1.665	−0.911	−0.911	0.335	−0.305	0.830
4	−0.364	0.366	0.366	1.636	0.599	0.134
5	−0.081	−0.639	−0.639	1.919	−1.226	0.408
6	−0.735	−1.086	−1.086	1.265	−1.373	1.179
7	−2.507	0.67	0.670	−0.507	−0.340	0.449
8	−1.144	0.238	0.238	0.856	0.204	0.057
9	−2.419	−0.055	−0.055	−0.419	0.023	0.003
10	−3.151	1.148	1.148	−1.151	−1.322	1.318
11	−2.085	−0.135	−0.135	−0.085	0.012	0.018
12	−2.972	−0.299	−0.299	−0.972	0.291	0.089
13	−0.633	−1.027	−1.027	1.367	−1.404	1.055
14	−2.678	0.331	0.331	−0.678	−0.225	0.110
15	−7.095	2.622	2.622	−5.095	−13.360	6.875
Average	**−2.000**	**0.000**			**−1.381**	**0.934**

$$\hat{\beta_2} = \frac{-1.381}{0.934} = -1.479$$
$$\hat{\alpha} = \overline{Y} - \hat{\beta_2}\overline{X_2} = -2.000$$

So:

$$Y = -2.000 - 1.479X_2$$

c. The key statement from the chapter for this question is:

"the estimator $\hat{\beta}_1$ converges to $\beta_1 + \beta_2\delta$"

So, part a gives that:

$$\beta_1 + \beta_2\delta_1 = 0$$

And part b gives that:

$$\beta_2 + \beta_1\delta_2 = -1.479$$

	X1	X2	$(X_1-\overline{X_1})$	$(X_2-\overline{X_2})$	$(X_1-\overline{X_1})(X_2-\overline{X_2})$	$(X_1-\overline{X_1})^2$	$(X_2-\overline{X_2})^2$
1	−0.409	−0.008	−0.409	−0.008	0.003	0.167	0.000
2	0.397	−1.216	0.397	−1.216	−0.483	0.158	1.478
3	−0.856	−0.911	−0.856	−0.911	0.780	0.733	0.830
4	1.682	0.366	1.682	0.366	0.616	2.829	0.134
5	0.455	−0.639	0.455	−0.639	-0.291	0.207	0.408
6	−1.39	−1.086	−1.390	−1.086	1.510	1.932	1.179
7	0.954	0.67	0.954	0.670	0.639	0.910	0.449
8	1.021	0.238	1.021	0.238	0.243	1.042	0.057
9	−0.156	−0.055	−0.156	−0.055	0.009	0.024	0.003
10	1.382	1.148	1.382	1.148	1.587	1.910	1.318
11	−0.562	−0.135	−0.562	−0.135	0.076	0.316	0.018
12	−1.554	−0.299	−1.554	−0.299	0.465	2.415	0.089
13	−1.123	−1.027	−1.123	−1.027	1.153	1.261	1.055
14	−0.124	0.331	−0.124	0.331	−0.041	0.015	0.110
15	0.284	2.622	0.284	2.622	0.744	0.081	6.875
Average	**0.000**	**0.000**			**0.467**	**0.933**	**0.934**

$$\delta_1 = \frac{Cov[X_1,X_2]}{V[X_1]} = \frac{0.467}{0.933} = 0.501$$

and:

$$\delta_2 = \frac{Cov[X_1,X_2]}{V[X_2]} = \frac{0.467}{0.934} = 0.500$$

Note: In this case these quantities are essentially equal, but that will not usually be the case.

Plugging back in the 2×2 system of equations:

$$\beta_1 + 0.501\beta_2 = 0$$
$$0.500\beta_1 + \beta_2 = -1.479$$

Solving yields that

$$\beta_1 = 0.989 \quad \beta_2 = -1.9733$$

And plugging these back into the equation to find α:

$$\overline{Y} = \alpha + \beta_1\overline{X_1} + \beta_2\overline{X_2} = -2.000$$
$$Y = -2.000 + 0.989X_1 - 1.9733X_2$$

d. All approaches should provide the same answer.

9.12 The first task is to calculate the squared residuals:

	M1	M2	M3	M1	M2	M3
1	2.1	0.6	1.1	4.41	0.36	1.21
2	−0.1	0.4	2.5	0.01	0.16	6.25
3	−0.8	0.1	1	0.64	0.01	1
4	−2.6	0.5	−0.1	6.76	0.25	0.01
5	−1.1	−0.6	1.3	1.21	0.36	1.69
6	-0.4	−0.9	0.5	0.16	0.81	0.25
7	0.6	0	0	0.36	0	0
8	0.9	−1	−0.4	0.81	1	0.16
9	1.2	0.8	−0.1	1.44	0.64	0.01
			TOTAL RSS	15.8	3.59	10.58
			CV RSS	5.06	1.42	0.17

The model selected is the one that has the smallest RSS within the blue out-of-sample boxes—this is M3.

9.13 **a.** Following the methodology of the first example:

	Alpha	Beta
Entire	3.26	3.49
First nine	0.10	2.96

Completing the table:

	Y	X	Y1	Y2	(Y1−Y2)^2	Y1−Y
1	−10.42	−5	−14.181	−14.707	0.277	−3.761
2	−14.12	−4	−10.692	−11.745	1.107	3.428
3	−14.72	−3	−7.204	−8.783	2.491	7.516
4	−8.25	−2	−3.716	−5.821	4.428	4.534
5	2.3	−1	−0.228	−2.858	6.919	−2.528
6	1.59	0	3.260	0.104	9.963	1.670
7	5.87	2	10.236	6.028	17.713	4.366
8	14.13	3	13.724	8.990	22.418	−0.406
9	6.78	4	17.213	11.952	27.676	10.433
10	32	1	6.748	3.066	13.561	−25.252
				Sum	106.552	
					SumSq/10	87.783

Y1 is the estimate using the entire data set

Y2 is the estimate using the first nine observations

So the Cook's distance for the last data point is

$$D = \frac{106.552}{2 * 87.783} = 0.61$$

b. The last data point does not have a major influence on the OLS parameters. However, it may still be an outlier, and other tests would need to be conducted to draw such a conclusion.

10

Stationary Time Series

Learning Objectives

After completing this reading you should be able to:

- Describe the requirements for a series to be covariance-stationary.
- Define the autocovariance function and the autocorrelation function.
- Define white noise, and describe independent white noise and normal (Gaussian) white noise.
- Define and describe the properties of autoregressive (AR) processes.
- Define and describe the properties of moving average (MA) processes.
- Explain how a lag operator works.
- Explain mean reversion and calculate a mean-reverting level.
- Define and describe the properties of autoregressive moving average (ARMA) processes.
- Describe the application of AR, MA, and ARMA processes.
- Describe sample autocorrelation and partial autocorrelation.
- Describe the Box-Pierce Q statistic and the Ljung-Box Q statistic.
- Explain how forecasts are generated from ARMA models.
- Describe the role of mean reversion in long-horizon forecasts.
- Explain how seasonality is modeled in a covariance-stationary ARMA.

Time-series analysis is a fundamental tool in finance and risk management. Many key time series (e.g., interest rates and spreads) have predictable components and building accurate time-series models allows past values to be used to forecast future changes in these series.

A time series can be decomposed into three distinct components: the trend, which captures the changes in the level of the time series over time; the seasonal component, which captures predictable changes in the time series according to the time of year; and the cyclical component, which captures the cycles in the data. Whereas the first two components are deterministic, the third component is determined by both the shocks to the process and the memory (i.e., persistence) of the process.

This chapter focuses on the cyclical component. The properties of this component determine whether past deviations from trends or seasonal components are useful for forecasting the future. The next chapter extends the model to include trends and deterministic seasonal components and examines methods that can be used to remove trends.

This chapter begins by introducing time-series and linear processes. This class of processes is broad and includes most common time-series models. It then introduces a key concept in time-series analysis called covariance stationarity. A time series is covariance-stationary if its first two moments do not change over time. Importantly, any time series that is covariance-stationary can be described by a linear process.

While linear processes are very general, they are also not directly applicable to modeling. Instead, two classes of models are used to approximate general linear processes: autoregressions (AR) and moving averages (MAs). The properties of these processes are exploited when modeling data by comparing the theoretical structure of the different model classes to sample statistics. These inspections serve as a guide when selecting a model. Model selection is refined using information criteria (IC) that formalize the tradeoff between bias and variance when choosing a specification. This chapter covers the two most widely used criteria.

The chapter concludes by examining how these models are used to produce out-of-sample forecasts and how ARs and MAs can be adapted to time series with seasonal dynamics.

10.1 STOCHASTIC PROCESSES

Stochastic processes are ordered collections of random variables. They are denoted using $\{Y_t\}$, reflecting the fact that they are sequences of random variables that are ordered in time (t) (i.e., so that Y_s is observed before Y_t whenever $s < t$). The ordering of a time series is important when predicting future values using past observations.

Stochastic processes can describe a wide variety of data-generating models. One of the simplest useful stochastic processes is $Y_t \stackrel{iid}{\sim} N(\mu, \sigma^2)$, which is a reasonable description of the data-generating process for returns on some financial assets. In this process, Y_t has no deterministic structure and is pure noise. This provides the building block for other, more sophisticated models.

A leading example of a more complex stochastic process (and one that is a focus of this chapter) is a first-order autoregression, also known as an AR(1):

$$Y_t = \delta + \phi Y_{t-1} + \epsilon_t, \qquad (10.1)$$

where δ is a constant, ϕ is a model parameter measuring the strength of the relationship between two consecutive observations, and ϵ_t is a shock.

This first-order AR process can describe a wide variety of financial and economic time series (e.g., interest rates, commodity convenience yields, or the growth rate of industrial production).

Figure 10.1 shows examples of time series that are produced by stochastic processes with very different properties. The top-left panel shows the monthly return on the S&P 500 index. This time series appears to be mostly noise and so is hard to predict. The top-right panel contains the VIX index, which appears to be very slowly mean-reverting (i.e., it tends to gradually go back to its average level so that deviations from the long-run average are long-lived but temporary).

The bottom panels show two important measures of interest rates: the slope and curvature of the US Treasury yield curve. The slope is defined as the difference between the yield on a ten-year bond and that on a one-year bond. The curvature is defined as the difference between two slopes:

$$(Y_{10} - Y_5) - (Y_5 - Y_1),$$

where Y_m is the yield on a bond with a maturity of m years. Both series are mean-reverting and have important cyclical components that are related to the state of the economy.

This chapter focuses exclusively on linear process because a wide range of processes, including nonlinear processes, have linear representations.

Any linear process can be written as:

$$\begin{aligned} Y_t &= \delta_t + \psi_0\epsilon_t + \psi_1\epsilon_{t-1} + \psi_2\epsilon_{t-2} + \cdots \\ &= \delta_t + \sum_{i=0}^{\infty}\psi_i\epsilon_{t-i} \end{aligned} \qquad (10.2)$$

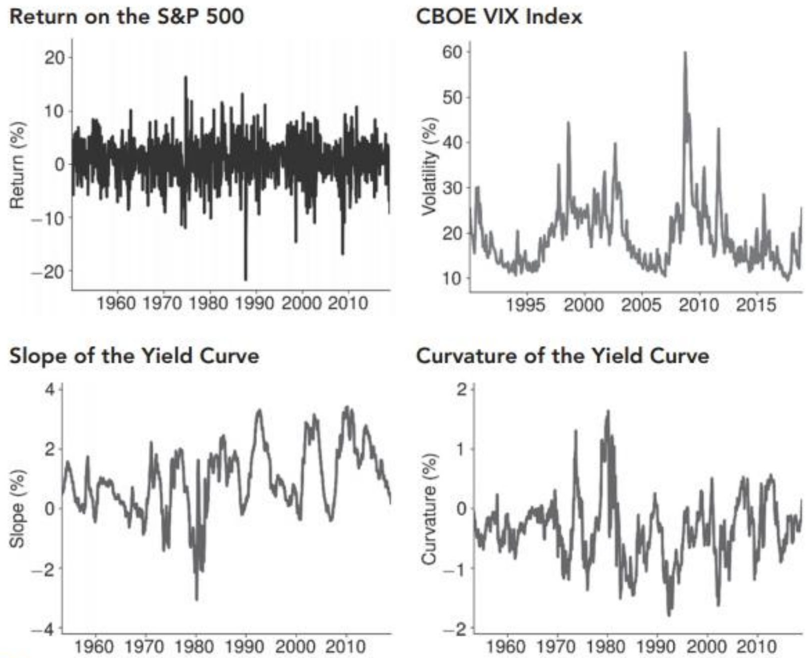


Figure 10.1 **These plots show examples of time series. The top-left panel shows the monthly return on the S&P 500. The top-right panel shows the end-of-month value of the VIX index. The bottom panels show the slope and curvature of the US Treasury yield curve.**

This process is linear in $\{\epsilon_t\}$, which is a mean zero stochastic process commonly referred to as the shock. In addition, the process δ_t is deterministic and the coefficients on the shocks, ψ_i, are constant.[1]

Linear processes can be directly related to linear models, which are the workhorses of time-series analysis and the most widely used class of models for modeling and forecasting the conditional mean of a process. They can be defined as:

$$E[Y_{t+1}|\mathcal{F}_t]$$

where $\mathcal{F}_t$ is known as the information set and includes everything known at time t, including the history of Y (Y_t, Y_{t-1}, Y_{t-2}, ...).

[1] This chapter only studies models where $\delta_t = \delta$ for all t, so that the deterministic process is constant. Chapter 11 relaxes this restriction to include models where δ_t explicitly depends on t to account for both time trends and seasonal effects.

10.2 COVARIANCE STATIONARITY

Stationarity is a key concept that formalizes the structure of a time series and justifies the use of historical data to build models. Covariance stationarity depends on the first two moments of a time series: the mean and the autocovariances.

Autocovariance is a time-series specific concept. The autocovariance is defined as the covariance between a stochastic process at different points in time. Its definition is the time-series analog of the covariance between two random variables. The h^{th} autocovariance is defined as:

$$\gamma_{t,h} = E[(Y_t - E[Y_t])(Y_{t-h} - E[Y_{t-h}])] \quad (10.3)$$

Autocovariance is denoted using γ, where the subscripts denote the period (i.e., t) and the lag (i.e., h) between observations. When $h = 0$ then:

$$\gamma_{t,0} = E[(Y_t - E[Y_t])^2],$$

which is the variance of Y_t.

This new measure is needed to define the properties of a covariance-stationary time series. A time series is covariance-stationary if its first two moments satisfy three key properties.

1. The mean is constant and does not change over time (i.e., $E[Y_t] = \mu$ for all t).
2. The variance is finite and does not change over time (i.e., $V[Y_t] = \gamma_0 < \infty$ for all t).
3. The autocovariance is finite, does not change over time, and only depends on the distance between observation h (i.e., $Cov[Y_t, Y_{t-h}] = \gamma_h$ for all t).

Covariance stationarity is important when modeling and forecasting time series.

- A covariance-stationary time series has constant relationships over time.
- Parameters estimated from non-stationary time series are more difficult to interpret.

A constant relationship over time allows historical data to be used to estimate models that are applicable to future, out-of-sample observations. Meanwhile, non-stationary time series can be difficult to interpret because their estimated parameters are not asymptotically normally distributed. Furthermore, non-stationary time series are subject to spurious relationships where unrelated series appear to have strong, statistically significant correlations.

When $\{Y_t\}$ is covariance-stationary (i.e., the autocovariance does not depend on time), the autocorrelation at lag h is defined as the ratio:

$$\rho_h = \frac{Cov[Y_t, Y_{t-h}]}{\sqrt{V[Y_t]V[Y_{t-h}]}} = \frac{\gamma_h}{\sqrt{\gamma_0\gamma_0}} = \frac{\gamma_h}{\gamma_0} \qquad (10.4)$$

Autocorrelations, like correlations, are always between −1 and 1, inclusive. The set of autocovariances for lags $h = \pm 1, \pm 2, \ldots,$ is known as the autocovariance function, which returns the autocovariance between Y_t and Y_{t-h}:

$$\gamma(h) = \gamma_{|h|} \qquad (10.5)$$

This function is only well defined for covariance-stationary processes. The symmetry follows from the third property of covariance stationarity, because the autocovariance depends only on h and not t, so that:

$$Cov[Y_t, Y_{t-h}] = Cov[Y_{t+h}, Y_t]$$

The autocorrelation function (ACF) is similarly defined using the autocorrelations:

$$\rho(h) = \rho_{|h|} \qquad (10.6)$$

The ACF is typically paired with a closely related measure, called the partial autocorrelation function (PACF). The partial autocorrelation between Y_t and Y_{t-h} measures the strength of the correlation between these two values after *controlling* for the values between them (i.e., the intermediate lags $Y_{t-1}, Y_{t-2}, \ldots Y_{t-h+1}$).

The partial autocorrelation is a nonlinear transformation of the ACF. It is identical to the autocorrelation at the first lag because there are no values between Y_t and Y_{t-1}, but it differs at all other lags. It is common to use the notation $\alpha(h)$ to represent this function. The ACF and PACF are widely used in model selection since each linear time series model has a characteristic shape for each of these functions.

10.3 WHITE NOISE

White noise is the fundamental building block of any time-series model. A white noise process is denoted:

$$\epsilon_t \sim WN(0, \sigma^2)$$

where σ^2 is the variance of the shock. Note that any white noise process is covariance-stationary because the first two moments are time-invariant, and the variance is finite.

Shocks from a white noise process are used to simulate data. White noise processes $\{\epsilon_t\}$ have three properties.

1. Mean zero (i.e., $E[\epsilon_t] = 0$). This property is a convenience, because any process with an error that has non-zero mean can always be defined in terms of a mean-zero error.
2. Constant and finite variance (i.e., $V[\epsilon_t] = \sigma^2 < \infty$). This is a technical assumption that is needed for the third property.
3. Zero autocorrelation and autocovariance (i.e., $Cov[\epsilon_t, \epsilon_{t-h}] = E[\epsilon_t\epsilon_{t-h}] = 0$ for all $h \neq 0$).

The lack of correlation is the essential characteristic of a white noise process and plays a key role in the estimation of time-series model parameters. Maintaining this assumption forces all autocorrelation in a time series to be driven by model parameters and not the relationships between the shocks. Testing whether estimated models produce shocks that are consistent with this property is a critical step in validating the specification of a model.

Note that any distribution having a finite variance and non-zero mean can be transformed into a white noise process by subtracting the mean. For example, if $\eta_t \overset{iid}{\sim} \chi^2_\nu$, then $\epsilon_t = \eta_t - \nu$ has mean zero and finite variance, so is a white noise process (that has a positive skew).

Independent, identically distributed random variables are a simple but important special case of white noise processes. Any iid

sequence that has mean zero and finite variance is white noise [e.g., $\epsilon_t \overset{iid}{\sim} G(0, \sigma^2)$, where $G(0, \sigma^2)$ is a distribution function[2] of a random variable that has mean zero and variance σ^2]. Gaussian white noise is a special case of iid noise, where $\epsilon_t \overset{iid}{\sim} N(0, \sigma^2)$.

White noise processes do not require a specific distributional assumption, and so assuming that model shocks are normally distributed is a stronger assumption than is required in time-series analysis. While convenient, this assumption is also often empirically refuted by financial asset data.

Dependent White Noise

White noise is uncorrelated over time but is not necessarily independent. Dependent white noise is particularly important in finance and risk management because asset returns are unpredictable but have persistent time-varying volatility. Such a process is linearly independent but nonlinearly dependent.

Dependent white noise relaxes the iid assumption while maintaining the three properties of white noise. A leading example of a dependent white noise process is known as an Autoregressive Conditional Heteroskedasticity (ARCH) process. The variance of a shock from an ARCH process depends on the magnitude of the previous shock. This process exhibits a property called volatility clustering, where volatility can be above or below its long-run level for many consecutive periods. Volatility clustering is an important feature of many financial time series, especially asset returns. The dependence in an ARCH process leads it to have a predictable variance but not a predictable mean, and shocks from an ARCH process are not correlated over time.

Wold's Theorem

Wold's theorem establishes the key role of white noise in any covariance-stationary process. It also provides an important justification for using linear processes to model covariance-stationary time series. If $\{Y_t\}$ is a mean-zero covariance-stationary process, then it can be expressed as:

$$\begin{aligned} Y_t &= \epsilon_t + \psi_1\epsilon_{t-1} + \psi_2\epsilon_{t-1} + \dots \\ &= \sum_{i=0}^{\infty}\psi_i\epsilon_{t-i,} \end{aligned} \qquad (10.7)$$

where $\psi_0, \psi_1, \dots$ are constants, $\{\epsilon_t\}$ is a white noise process, $\psi_0 = 1$, and $\sum_{j=0}^{\infty}\psi_i^2 < \infty$. Wold's theorem also states that this representation of a covariance-stationary process is unique.

[2] Note that G is not necessarily a normal distribution. For example, if G is a generalized Student's t_ν distribution with $\nu > 2$, then the shocks have heavy tails.

10.4 AUTOREGRESSIVE (AR) MODELS

Autoregressive models are the most widely applied time-series models in finance and economics. These models relate the current value of a stochastic process (i.e., Y_t) to its previous value (i.e., Y_{t-1}). Recall that a first order AR process, which can be denoted by AR(1), evolves according to:

$$Y_t = \delta + \phi Y_{t-1} + \epsilon_t,$$

where δ is called the intercept, ϕ is the AR parameter, and the shock $\epsilon_t \sim WN(0, \sigma^2)$.

The AR parameter determines the persistence of Y_t. This means that an AR(1) process is covariance-stationary when $|\phi| < 1$ and non-stationary when $\phi = 1$. This chapter focuses exclusively on the covariance-stationary case and leaves the treatment of non-stationary time series to the next chapter.

Because $\{Y_t\}$ is assumed to be covariance-stationary, the mean, variance, and autocovariances are all constant. Note that the mean of Y_t *is not* δ – it depends on both δ and ϕ.

Denoting the mean of Y_t by μ and using the property of a covariance-stationary time series that:

$$E[Y_t] = E[Y_{t-1}] = \mu$$

the long-run (or unconditional) mean is

$$\begin{aligned} E[Y_t] &= \delta + \phi E[Y_{t-1}] + E[\epsilon_t] \\ \mu &= \delta + \phi\mu + 0 \\ \mu(1-\phi) &= \delta \\ \mu &= \frac{\delta}{1-\phi} \end{aligned} \qquad (10.8)$$

Thus, the long-run mean depends on both parameters in the model.

The variance of Y_t, which is denoted by γ_0, also depends on ϕ:

$$\begin{aligned} V[Y_t] &= V[\delta + \phi Y_{t-1} + \epsilon_t] \\ &= \phi^2 V[Y_{t-1}] + V[\epsilon_t] + \text{Cov}[Y_{t-1}, \epsilon_t] \\ \gamma_0 &= \phi^2\gamma_0 + \sigma^2 + 0 \\ \gamma_0 &= \frac{\sigma^2}{1-\phi^2} \end{aligned} \qquad (10.9)$$

Specifically, the variance of Y_t depends on both the variance of the shocks (i.e., σ^2) and the AR parameter (i.e., ϕ). This derivation relies on the white noise assumption for ϵ_t so that $\text{Cov}[Y_{t-1}, \epsilon_t]$ is zero because Y_{t-1} depends on $\epsilon_{t-1}, \epsilon_{t-2}, \dots$

Autocovariances and Autocorrelations

The first autocovariance is in the AR(1) process is:

$$\begin{aligned}\text{Cov}[Y_t, Y_{t-1}] &= \text{Cov}[\delta + \phi Y_{t-1} + \epsilon_t, Y_{t-1}] \\ &= \phi\text{Cov}[Y_{t-1}, Y_{t-1}] + \text{Cov}[Y_{t-1}, \epsilon_t] \\ &= \phi\gamma_0 \qquad (10.10)\end{aligned}$$

The remaining autocovariances can be recursively computed by noting that:

$$\begin{aligned}\text{Cov}[Y_t, Y_{t-h}] &= \text{Cov}[\delta + \phi Y_{t-1} + \epsilon_t, Y_{t-h}] \\ &= \phi\text{Cov}[Y_{t-1}, Y_{t-h}] + \text{Cov}[Y_{t-h}, \epsilon_t] \\ &= \phi\gamma_{h-1} \qquad (10.11)\end{aligned}$$

Note that $\text{Cov}[Y_{t-h}, \epsilon_t] = 0$ because Y_{t-h} depends on $\epsilon_{t-1}, \epsilon_{t-2}, \ldots$, whereas ϵ_t happens *after* the final shock in Y_{t-h}. Applying the recursion, $\gamma_2 = \phi\gamma_1 = \phi^2\gamma_0$, and (in general):

$$\gamma_h = \phi^h\gamma_0$$

The autocovariance function is then:

$$\gamma(h) = \phi^{|h|}\gamma_0 \qquad (10.12)$$

and the ACF is

$$\rho(h) = \frac{\phi^h\gamma_0}{\gamma_0} = \phi^{|h|} \qquad (10.13)$$

Provided that $|\phi| < 1$, the ACF geometrically decays to zero as h increases. It also oscillates between negative and positive values if $-1 < \phi < 0$. In practice, negative values of ϕ are uncommon in economic and financial time series.

Finally, the PACF of an AR(1) is:

$$\alpha(h) = \begin{cases}\phi^{|h|} & h \in \{0, \pm 1\} \\ 0 & h \geq 2\end{cases} \qquad (10.14)$$

The PACF is non-zero only for the first lag. These general patterns—slow decay in the ACF and a steep cutoff in the PACF—play a key role when choosing appropriate models to apply to a data set.

Figure 10.2 contains the autocorrelation and PACFs of two AR(1) models. The left panel shows the ACF and PACF when $\phi = 0.7$. The ACF decays slowly to zero and the PACF is 0.7 at the first lag and then zero for all other lags. The right panel shows the ACF and PACF when ϕ is negative. The pattern is similar, with the additional feature that the ACF oscillates every other lag.

The Lag Operator

The lag operator is a convenient tool for expressing and manipulating more complex time-series models. The lag operator L shifts the time index of an observation, so that $LY_t = Y_{t-1}$. It has six properties.

1. When applied once, the lag operator shifts the time index back one observation, $LY_t = Y_{t-1}$
2. Applying the lag operator p times results in the p^{th} lag of a variable, $L^pY_t = Y_{t-p}$ (e.g., $L^2Y_t = L(LY_t) = LY_{t-1} = Y_{t-2}$)
3. The lag operator applied to a constant is just the constant (i.e., $L\delta = \delta$)
4. The p^{th} order lag polynomial is written:

   $$a(L) = 1 + a_1L + a_2L^2 + \ldots + a_pL^p$$

   For example:

   $$a(L)Y_t = Y_t + a_1Y_{t-1} + a_2Y_{t-2} + \ldots + a_pY_{t-p}$$

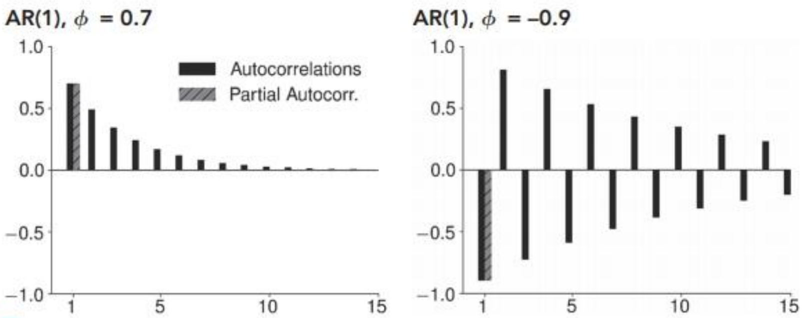


Figure 10.2 **Plot of the ACF and PACF for the first 15 lags from two AR(1) models.**

5. Lag polynomials can be multiplied. For example, if $a(L)$ and $b(L)$ are first-order polynomials, then:

$$a(L)b(L)Y_t = (1 + a_1L)(1 + b_1L)Y_t = (1 + a_1L)(Y_t + b_1Y_{t-1}) = Y_t + b_1Y_{t-1} + a_1Y_{t-1} + a_1b_1Y_{t-2}$$

Polynomial multiplication is commutative so that $a(L)b(L) = b(L)a(L)$.

6. If the coefficients in the lag polynomial satisfy some technical conditions, the polynomial can be inverted so that $a(L)a(L)^{-1} = 1$. When $a(L)$ is the first-order polynomial $1 - a_1L$, it is invertible if $|a_1| < 1$ and its inverse is

$$(1 - aL)^{-1} = \sum_{i=0}^{\infty} a^i L^i = 1 + aL + a^2L^2 + a^2L^3 + \dots \quad (10.15)$$

While it is usually not necessary to manually invert a lag polynomial, the concept of invertibility is useful for two reasons. First, an AR process is only covariance-stationary if its lag polynomial is invertible. For example, the AR(1):

$$Y_t = \delta + \phi Y_{t-1} + \epsilon_t$$

can be expressed with a lag polynomial:

$$Y_t = \delta + \phi L Y_t + \epsilon_t$$
$$(1 - \phi L)Y_t = \delta + \epsilon_t$$

If $|\phi| < 1$ then this lag polynomial is invertible:

$$(1 - \phi L)^{-1}(1 - \phi L)Y_t = (1 - \phi L)^{-1}\delta + (1 - \phi L)^{-1}\epsilon_t$$
$$\Rightarrow Y_t = \delta \sum_{i=0}^{\infty} \phi^i + \sum_{j=0}^{\infty} \phi^j L^j \epsilon_t = \frac{\delta}{1 - \phi} + \sum_{i=0}^{\infty} \phi^i \epsilon_{t-i}$$

Second, invertibility plays a key role when selecting a unique model for a time series using the Box-Jenkins methodology. Model building and the role of invertibility is covered in more detail in Section 10.8.

For higher order polynomials, it is difficult to define the values of the model coefficients where the polynomial is invertible—and hence the parameter values where an AR is covariance-stationary. In such cases, the use of characteristic equations becomes necessary. This method is explained in more detail in the Appendix to this chapter.

AR(*p*)

The p^{th} order AR process generalizes the first-order process to include p lags of Y in the model. The AR(p) model is:

$$Y_t = \delta + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p} + \epsilon_t \quad (10.16)$$

Deriving the properties for an AR(p) model is more involved than for an AR(1). Instead, this chapter focuses on how the properties of an AR(p) broadly mirror those of an AR(1).

Note that AR(p) model can also be written using a lag polynomial:

$$(1 - \phi_1 L - \dots - \phi_p L^p)Y_t = \delta + \epsilon_t$$

When $\{Y_t\}$ is covariance-stationary, the long-run mean is:[3]

$$E[Y_t] = \mu = \frac{\delta}{1 - \phi_1 - \phi_2 - \dots - \phi_p} \quad (10.17)$$

and the long-run variance is

$$V[Y_t] = \gamma_0 = \frac{\sigma^2}{1 - \phi_1 - \phi_2 - \dots - \phi_p}$$

This point acts as an attractor, in the sense that an AR process tends to move closer to the mean. Figure 10.3 shows the sample paths of two persistent AR processes. The first is an AR(1) with an AR coefficient of 0.97. This process is very persistent, although it is mean reverting to μ. The second process is $Y_t = 1.6Y_{t-1} - 0.7Y_{t-2} + \epsilon_t$. This process is less persistent than the AR(1) and crosses the mean more frequently.[4]

Note that expression for the long-run mean suggests that $\phi_1 + \dots + \phi_p < 1$ is a necessary but not sufficient condition for stationarity. The condition is simple to check, and if the sum of the coefficients is equal to or larger than 1, then the process cannot be covariance-stationary.[5]

The autocorrelation and PACFs of an AR(p) share a common structure with the ACF and PACF of an AR(1). For example, the ACF of an AR(p) also decays to zero as the lag length increases and may oscillate. However, higher-order ARs can produce more complex patterns in their ACFs than an AR(1).

Meanwhile, the PACF of an AR(p) shows a sharp drop-off at p lags. This pattern naturally generalizes the shape of the PACF in the AR(1), which cuts-off at one lag.

[3] This model is covariance-stationary if the roots of the associated characteristic polynomial are all greater than 1 in absolute value, in which case they are said to 'lie outside the unit circle.' See the Appendix for further details and an example.

[4] A useful measure of persistence in an AR(p) is the largest absolute root in the characteristic polynomial. The roots in this equation are complex and the absolute value (modulus) of the largest root is 1.119.

[5] Although note that even if the condition is satisfied, the process may still not be stationary.

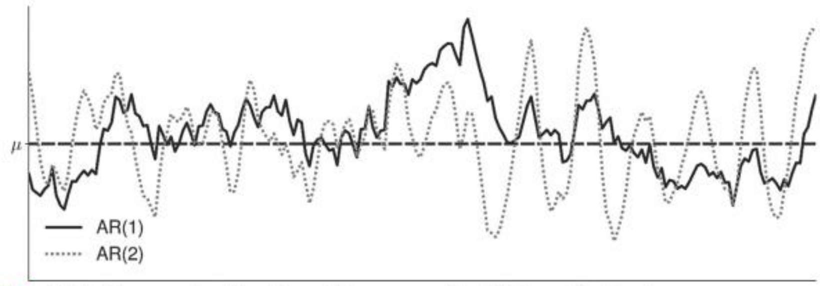


Figure 10.3 Plot of the sample paths of two AR processes. The AR(1) specification is $Y_t = 0.97Y_{t-1} + \epsilon_t$ and the AR(2) specification is $Y_t = 1.6Y_{t-1} - 0.8Y_{t-2} + \epsilon_t$. Both simulated paths use the same white noise innovations.

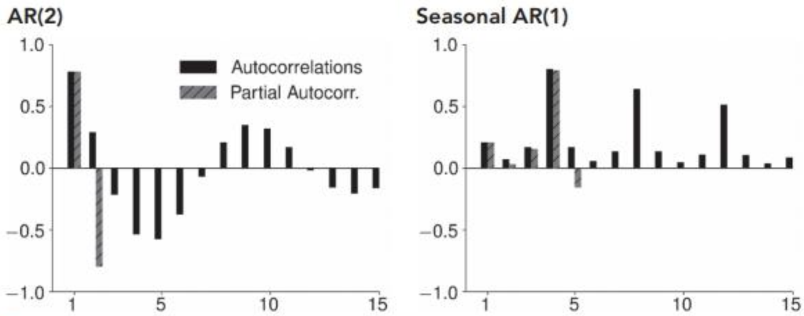


Figure 10.4 Plots of the autocorrelation and PACFs for an AR(2) (left panel) and a Seasonal AR(1) (right panel) for quarterly data.

The left panel of Figure 10.4 shows the first 15 values for the ACF and PACF of the AR(2):

$$Y_t = \delta + 1.4Y_{t-1} - 0.8Y_{t-2} + \epsilon_t$$

The ACF oscillates and decays slowly to zero as the lag length increases. Unlike the AR(1) with a negative coefficient, the oscillation does not occur in every period. Instead, it oscillates every four or five periods, which produces persistent cycles in the time series.

The right panel shows the ACF and PACF of a model known as a Seasonal AR. This model is easiest to express using the lag polynomial form:

$$(1 - \phi_4 L^4)(1 - \phi_1 L)Y_t = \epsilon_t$$

This model has a quarterly seasonality where the shock lagged four quarters has a distinct effect on the current value of Y_t. This model is technically an AR(5) because $(1 - \phi_4 L^4)(1 - \phi_1 L) = (1 - \phi_1 L - \phi_4 L^4 + \phi_1\phi_4 L^5)$ and:

$$(1 - \phi_1 L - \phi_4 L^4 + \phi_1\phi_4 L^5)Y_t = \delta + \epsilon_t$$
$$Y_t = \delta + \phi_1 Y_{t-1} + \phi_4 Y_{t-4} - \phi_1\phi_4 Y_{t-5} + \epsilon_t$$

The ACF and the PACF are consistent with the expanded model.

Suppose now that the key coefficients in this model take the values $\phi_1 = 0.2$ and $\phi_4 = 0.8$, which would indicate a small amount of short-run persistence coupled with a large amount of seasonal persistence. The seasonality produces the large spikes in the ACF every four lags. As expected, the PACF is non-zero for the first five lags because this model is effectively an AR(5).

YULE-WALKER EQUATIONS

The Yule-Walker (YW) equations provide a set of expression that relate the parameters of an AR to the autocovariances of the AR process. This approach uses $p + 1$ equations to solve for the long-run variance γ_0 and the first p autocorrelations. Autocovariances (or autocorrelations) at lags larger than p are then easily computed with a recursive structure starting from the first p autocovariances.

The equations are:

$$\text{Cov}[Y_t, Y_t] = \text{Cov}[\delta + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \epsilon_t, Y_t]$$
$$\text{Cov}[Y_t, Y_{t-1}] = \text{Cov}[\delta + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \epsilon_t, Y_{t-1}]$$
$$\vdots$$
$$\text{Cov}[Y_t, Y_{t-p}] = \text{Cov}[\delta + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \epsilon_t, Y_{t-p}]$$

These equations appear simpler when expressed using the standard autocovariance notation γ_j:

$$\gamma_0 = \phi_1\gamma_1 + \phi_2\gamma_2 + \cdots + \phi_p\gamma_p + \sigma^2$$
$$\gamma_1 = \phi_1\gamma_0 + \phi_2\gamma_1 + \cdots + \phi_p\gamma_{p-1}$$
$$\vdots$$
$$\gamma_p = \phi_1\gamma_{p-1} + \phi_2\gamma_{p-2} + \cdots + \phi_p\gamma_0$$

Excluding the first equation, dividing each row by the long-run variance γ_0 produces a set of equations that relate the autocorrelations:

$$\rho_1 = \phi_1 + \phi_2\rho_1 + \cdots + \phi_p\rho_{p-1}\vdots$$
$$\rho_p = \phi_1\rho_{p-1} + \phi_2\rho_{p-2} + \cdots + \phi_p \quad (10.18)$$

While this system of p equations has p unknown values, it always has a unique solution when $\phi_1, \ldots, \phi_p$ are compatible with an assumption of stationarity. The long-run variance can be computed using the first equation because $\gamma_j = \rho_j\gamma_0$, so that:

$$\gamma_0 = \phi_1\rho_1\gamma_0 + \phi_2\rho_2\gamma_0 + \cdots + \phi_p\rho_p\gamma_0 + \sigma^2$$
$$\gamma_0 = \frac{\sigma^2}{1 - \phi_1\rho_1 - \phi_2\rho_2 - \cdots - \phi_p\rho_p} \quad (10.19)$$

Higher-order autocorrelations are recursively computed using the relationship:

$$\rho_j = \phi_1\rho_{j-1} + \phi_2\rho_{j-2} + \cdots + \phi_p\rho_{j-p} \quad (10.20)$$

Applying the Yule-Walker Equations to an AR(2)

For example, suppose that:

$$Y_t = 1.4Y_{t-1} - 0.45Y_{t-2}$$

This AR(2) is stationary because the roots of the characteristic polynomial are 0.9 and 0.5. The YW equations are:

$$\rho_1 = 1.4 - 0.45\rho_1$$
$$\rho_2 = 1.4\rho_1 - 0.45$$

The solutions to these are:

$\rho_1 = 1.4/1.45 = 0.966$ and $\rho_2 = 1.4^2/1.45 - 0.45 = 0.902$

The remaining autocorrelations are recursively computed using:

$$\rho_j = 1.4\rho_{j-2} - 0.45\rho_{j-1}$$

Using this equation, next three autocorrelations can be computed as 0.828, 0.753, and 0.682.

Finally, the long-run variance is:

$$\gamma_0 = \frac{\sigma^2}{1 - 1.4 \times \frac{1.4}{1.45} + 0.45 \times \left(\frac{1.4^2}{1.45} - 0.45\right)} = \frac{\sigma^2}{0.054}$$

MODELING TIME-SERIES DATA: AN EXAMPLE

The concepts of this chapter are applied to modeling two important macrofinancial time series. The first is the default premium, defined as the difference in the interest rates of two corporate bond portfolios (i.e., a "safe" portfolio and a "risky" one). The safe portfolio holds only Aaa-rated securities and defaults among these bonds are extremely rare. Meanwhile, the risky portfolio holds Baa-rated corporate bonds, which is the lowest rating for investment grade securities.

The default premium is defined as the difference between the yields of these two portfolios:

$$DEF_t = Aaa_t - Baa_t$$

Observations of this series are available monthly from 1979 until the end of 2018.

The second series is the annualized growth rate of real Gross Domestic Product (real GDP) in the United States. Real GDP is a broad measure of economic activity that includes consumption and investment. Changes in price levels over time have been removed so that the growth rate only measures changes in economic activity (i.e., it is in real terms). Real GDP is measured quarterly, and so the annualized percentage growth rate is defined as:

$$RGDPG_t = 400 \times (\ln RGDP_t - \ln RGDP_{t-1}),$$

where $RGDP_t$ is the level of real GDP in period t. Data are available quarterly between 1947Q1 and 2018Q4.

Figure 10.5 contains plots of these two series. The left panel shows the default premium, and the right panel contains the growth rate of real GDP. The dark bands indicate recessions, as determined by the National Bureau of Economic Research.

Note that both series are responsive to the state of the economy: The default premium rises when economic conditions are poor, and so is countercyclical, whereas the real GDP growth rate is pro-cyclical by definition.

Results from fitting AR models with orders 1, 2, and 3 are reported in Table 10.1, and *t*-statistics are reported below each estimated coefficient. The models for the default premium (i.e., top panel) indicate that the series is highly persistent, with the AR(1) model having an estimated AR coefficient of 0.962. However, the AR(2) and AR(3) specifications have more complicated dynamics than the AR(1), and so their parameters cannot be as easily interpreted. Instead, the largest absolute roots from each model's characteristic equation (which are close to one) indicates the persistence in higher-order AR processes. These values are similar across the three specifications.

The three AR specifications estimated on the real GDP growth data indicate that the persistence for this data is lower than for that for the default premium.

The R^2 in a time-series model has the same interpretation as in a linear regression: it measures the proportion of the variation in Y_t that is explained by the model. The R^2 in all three real GDP growth models is also lower, indicating that it is more difficult to forecast GDP growth using historical data.

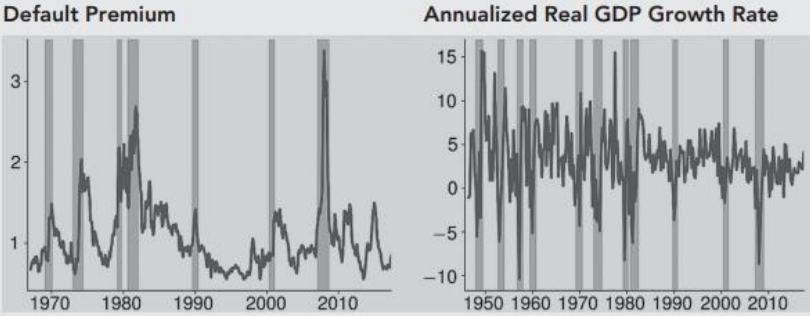


Figure 10.5 The left panel shows the default premium defined as the difference between the average interest rates on portfolios of Aaa- and Baa-rated bonds. The right panel shows the annualized growth rate of real GDP. Dark bars indicate recessions.

Table 10.1 Parameter Estimates from AR Models with Orders 1, 2, and 3. The Top Panel Reports Parameter Estimates Using the Default Premium, and the Bottom Panel Reports Parameter Estimates from Models Using Real GDP Growth. The Column Labeled R^2 Reports the Fit of the Model. The Final Column Reports the Inverses of the Absolute Values of the Roots of the Characteristic Equations (i.e. One Over the Roots, so that in the First Row, for Example, the Actual Root of the Characteristic Equation is 1/0.962 = 1.040, etc.)

Default Premium

δ	φ_1	φ_2	φ_3	R^2	Abs. Roots
0.040	0.962			0.927	0.962
(8.721)	(90.039)				
0.053	1.272	−0.322		0.935	0.924, 0.348
(11.897)	(33.294)	(−8.421)			
0.046	1.315	−0.490	0.132	0.936	0.944, 0.374, 0.374
(10.498)	(32.853)	(−7.716)	3.300		

Real GDP Growth Rate					
δ	φ_1	φ_2	φ_3	R^2	Abs. Roots
1.996	0.360			0.130	0.360
(9.658)	(6.544)				
1.765	0.319	0.114		0.141	0.533, 0.213
(8.609)	(5.455)	(1.938)			
1.963	0.333	0.150	−0.113	0.152	0.486, 0.486, 0.477
9.622	(5.681)	(2.448)	(−1.921)		

10.5 MOVING AVERAGE (MA) MODELS

A first-order MA, denoted MA(1), is defined as:

$$Y_t = \mu + \theta\epsilon_{t-1} + \epsilon_t, \quad (10.21)$$

where $\epsilon_t \sim WN(0, \sigma^2)$ is a white noise process.

The observed value of Y_t depends on both the contemporaneous shock ϵ_t and the previous shock. The parameter θ acts as a weight and determines the strength of the effect of the previous shock. The parameter μ is the mean of the process because:

$$\begin{aligned} E[Y_t] &= \mu + \theta E[\epsilon_{t-1}] + E[\epsilon_t] \\ &= \mu + \theta \times 0 + 0 \\ &= \mu \end{aligned} \quad (10.22)$$

When θ is positive, an MA(1) is persistent because two consecutive values are (on average) positively correlated. When θ is negative, the process aggressively mean reverts because the effect of the previous shock is reversed in the current period.

Moving averages are always covariance-stationary. An MA(1) has a limited memory, because only the shock in the previous period impacts the current value. The variance of an MA(1) is:

$$\begin{aligned} V[Y_t] &= V[\mu + \theta\epsilon_{t-1} + \epsilon_t] \\ &= \theta^2 V[\epsilon_{t-1}] + V[\epsilon_t] \\ &= (1 + \theta^2)\sigma^2 \end{aligned} \quad (10.23)$$

because the shocks are white noise and so are uncorrelated. Any MA(1) has exactly one non-zero autocorrelation, and the ACF for each lag length h is

$$\rho(h) = \begin{cases} 1 & h = 0 \\ \dfrac{\theta}{1+\theta^2} & h = 1 \\ 0 & h \geq 2 \end{cases} \quad (10.24)$$

The PACF of an MA(1), on the other hand, is complex and has non-zero values at all lags. This pattern is the inverse of what an AR(1) would produce.

The MA(1) generalizes to an MA(q), which includes q lags of the shock:

$$Y_t = \mu + \epsilon_t + \theta_1\epsilon_{t-1} + \cdots + \theta_q\epsilon_{t-q}. \quad (10.25)$$

Here, μ is the mean of Y_t because all shocks are white noise and so have zero expected value. The variance of an MA(q) can be shown to be

$$\gamma_0 = \sigma^2 (1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2). \quad (10.26)$$

The autocovariance function is

$$\gamma(h) = \begin{cases} \sigma^2 \sum_{i=0}^{q-h} \theta_i\theta_{i+h} & 0 \leq h \leq q \\ 0 & h > q \end{cases}, \quad (10.27)$$

where $\theta_0 = 1$.

The ACF is always zero for lags larger than q and the PACF is also complex. In general, the PACF of an MA(q) is non-zero at all lags.

Figure 10.6 plots the ACFs and PACFs for four MA processes. The top-left contains the values for an MA(1) with a positive coefficient. Here, the first autocorrelation is positive, and all other autocorrelations are zero. The PACF oscillates and decays slowly towards zero.

The top-right shows the ACF and PACF for an MA(1) with a large negative coefficient. The first autocorrelation is negative, and the remaining autocorrelations are all zero. The PACF decays slowly towards zero.

The bottom-left and right panels plot the values for an MA(2) and an MA(4), respectively. These both show the same pattern—one that is key to distinguishing MA models from

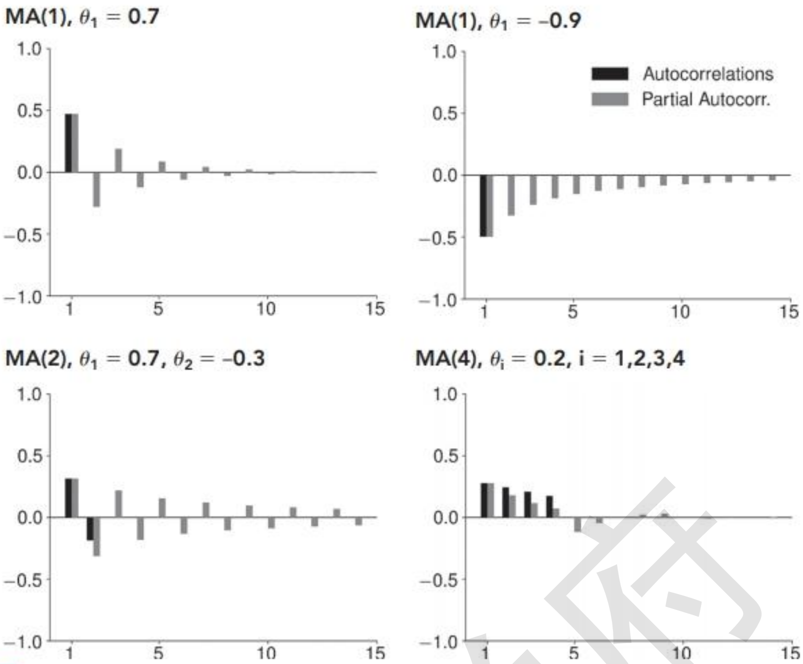


Figure 10.6 **These four panels show the ACFs and PACFs for two MA(1)s (top), an MA(2) (bottom-left) and an MA(4) (bottom-right). The ACF cuts off sharply at the MA order (*q*) while the PACF decays, possibly oscillating, over many lags.**

AR models: The ACF cuts off sharply at the order of the MA, whereas the PACF oscillates and decays towards zero over many lags. This is the opposite pattern from what is seen in Figure 10.2 and Figure 10.4 for the AR models, where the ACFs decayed slowly while the PACF cut off sharply.

When the stationarity condition applies, an AR(*p*) model can be written as an MA(∞). An AR(*q*) model is stationary if all the roots of its characteristic equation (see the Appendix) lie outside the unit circle. A property similar to stationarity applies to MA(*q*) models and is known as invertibility. An MA(*q*) model is invertible if all of the roots of its characteristic equation (defined in an identical fashion to that of an AR model) lie outside the unit circle. If an MA(*q*) model is invertible, it can be written as an AR(∞).

APPLYING MOVING AVERAGE MODELS: EXAMPLE CONTINUED

Moving average models are estimated for both the default premium and real GDP growth data. The estimated parameters, *t*-statistics and R^2 values of the estimated models are presented in Table 10.2.

While the default premium is a highly persistent series, an MA(*q*) is not an accurate approximation of its dynamics because these models are only autocorrelated with their first *q* lags. Note that the R^2 values of all models, even the MA(4), are lower than the R^2 in the AR(1) presented in Table 10.1. All estimated coefficients are statistically significant with large *t*-statistics.

GDP growth appears to be better described by a finite-order MA. While the R^2 from the MA(1) is less than that of the AR(1), the R^2 from the second and third-order models is similar. Adding a fourth MA lag does not change the model fit and $\hat{\theta}_4$ is not statistically different from zero.

Table 10.2 Parameter Estimates from MA Models with Orders between 1 and 4. The Top Panel Reports Parameter Estimates Using the Default Premium, and the Bottom Panel Reports Parameters Estimates from Models Using Real GDP Growth. The Column Labeled R^2 Reports the Fit of the Model

Default Premium

δ	θ_1	θ_2	θ_3	θ_4	R^2
1.074	0.915				0.688
(56.538)	(69.917)				
1.074	1.328	0.694			0.836
(49.449)	(48.559)	(27.838)			
1.073	1.432	1.221	0.616		0.892
(43.206)	(42.062)	(31.475)	(22.962)		
1.073	1.527	1.465	0.989	0.349	0.908
37.388	(39.277)	(25.735)	(20.561)	(10.396)	

Real GDP Growth Rate

δ	θ_1	θ_2	θ_3	θ_4	R^2
3.125	0.269				0.096
(11.686)	(5.725)				
3.117	0.307	0.227			0.146
(9.924)	(5.306)	(4.153)			
3.113	0.323	0.276	0.092		0.153
(9.031)	(5.576)	(4.364)	(1.548)		
3.113	0.328	0.279	0.096	0.014	0.153
(8.894)	(5.309)	(4.269)	1.531	(0.211)	

10.6 AUTOREGRESSIVE MOVING AVERAGE (ARMA) MODELS

Autoregressive Moving Average (ARMA) processes combine AR and MA processes. For example, a simple ARMA(1,1) evolves according to:

$$Y_t = \delta + \phi Y_{t-1} + \theta\epsilon_{t-1} + \epsilon_t \tag{10.28}$$

The mean of this process is

$$\mu = \delta/(1 - \phi)$$

The variance is

$$\gamma_0 = \sigma^2(1 + 2\phi\theta + \theta^2)/(1 - \phi^2)$$

The autocovariance function is

$$\gamma(h) = \begin{cases} \sigma^2 \dfrac{(1 + 2\phi\theta + \theta^2)}{1 - \phi^2} & h = 0 \\ \sigma^2 \dfrac{\phi(1 + \phi\theta) + \theta(1 + \phi\theta)}{1 - \phi^2} & h = 1 \\ \phi\gamma_{h-1} & h \geq 2 \end{cases} \tag{10.29}$$

The autocovariance function is complicated, even for an ARMA(1,1). The ACF decays as h increases and oscillates if $\phi < 0$. This is consistent with the ACF of an AR process. However, the PACF also slowly decays toward zero as the lag length increases. This behavior is consistent with a MA process. The slow decay of both the PACF and ACF is the key feature that distinguishes an ARMA from an AR or an MA process.

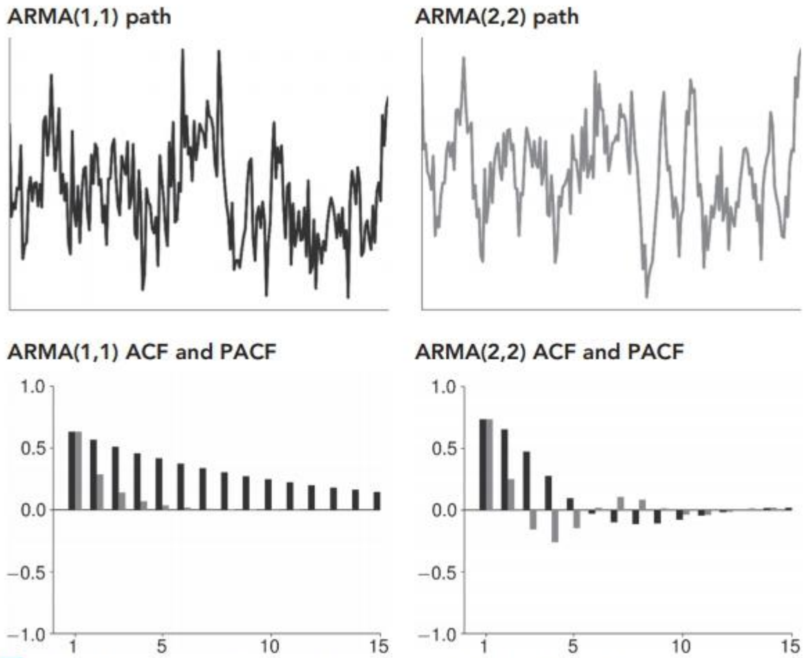


Figure 10.7 **The left panels plot a simulated path (top) and the ACF and PACF of the ARMA(1,1), $Y_t = 0.8Y_{t-1} - 0.4\epsilon_{t-1} + \epsilon_t$. The right panels plot a simulated path (top) and the ACF and PACF of the ARMA(2,2), $Y_t = 1.4Y_{t-1} - 0.6Y_{t-2} + 0.9\epsilon_{t-1} + 0.6\epsilon_{t-2} + \epsilon_t$.**

The left panels of Figure 10.7 show a simulated path (top) and the autocorrelation and PACFs (bottom) of an ARMA(1,1) with $\phi = 0.8$ and $\theta = -0.4$. Both functions decay slowly to zero as the lag length increases. The decay resembles that of an AR(1), although the level of the first autocorrelation is less than $\phi = 0.8$. This difference arises because the MA term eliminates some of the correlation with the first shock.

An ARMA(1,1) process is covariance-stationary if $|\phi| < 1$. The MA coefficient plays no role in determining whether the process is covariance-stationary, because any MA is covariance-stationary and the MA component only affects a single lag of the shock. The AR component, however, affects all lagged shocks and so if ϕ_1 is too large, then the time series is not covariance-stationary.

ARMA(p,q) processes combine AR(p) and MA(q) processes to produce models with more complicated dynamics. An ARMA(p,q) evolves according to

$$Y_t = \delta + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t \quad (10.30)$$

Or when compactly expressed using lag polynomials

$$(1 - \phi_1 L - \cdots - \phi_p L^p) Y_t = (1 + \theta_1 L + \cdots + \theta_q L^q)\epsilon_t$$

$$\phi(L) Y_t = \delta + \theta(L)\epsilon_t$$

Like an ARMA(1,1) model, the ARMA(p,q) is covariance-stationary if the AR component is covariance-stationary.[6]

The autocovariance and ACFs of ARMA processes are more complicated than in pure AR or MA models, although the general pattern in these is simple. ARMA(p,q) models have ACFs and PACFs that decay slowly to zero as the lag increases (while possibly oscillating).

[6] This requires the roots of the characteristic equation to all be greater than 1 in absolute value (see the Appendix).

APPLYING AUTOREGRESSIVE MOVING AVERAGE MODELS: EXAMPLE CONTINUED

Table 10.3 reports the parameter estimates, t-statistics and the model R^2 for a range of ARMA models fit to both the default premium and real GDP growth.

The default premium data are fitted using models with p or q up to 2. However, including a second lag of either the AR or MA term does not improve the model fit over the ARMA(1,1). Comparing the ARMA(1,1) with the AR(1) in Table 10.1, the AR parameter is virtually identical, and the MA parameter is positive. This configuration of parameters in the ARMA produces an ACF that lies below the implied ACF of the estimated AR model. Adding the MA component has improved the model fit as measured by the R^2 and the t-statistic on $\hat{\theta}_1$. The ARMA(1,1) fits slightly better than an AR(2) and about as well as an AR(3).

Models for real GDP growth also benefit from both terms. Adding the MA term allows the AR coefficients to increase, which provides more persistence while limiting the effect of the most recent shock. The final model, an ARMA(3,2), is selected using a model selection criterion (see Section 10.8). It does provide a better fit than any of the smaller models, although there is some risk of overfitting when using such a large model on quarterly data. The final column reports the inverses of the absolute value of the roots of the characteristic polynomial for the AR (i.e., one divided by the roots in each case). The maximum absolute root provides an indication of the persistence in the process. The AR(1) only has one root but the AR(2) has two. The largest root in both cases is $1/0.938 = 1.066$, which is close enough to one that the processes are stationary but highly persistent.

Table 10.3 **Parameter Estimates from ARMA Models. The Top Panel Reports Parameter Estimates Using the Default Premium, and the Bottom Panel Reports Parameter Estimates from Models on Real GDP Growth. The Column Labeled R^2 Reports the Fit of the Model. The Final Column Reports the Inverses of the Absolute Value of the Roots of the Characteristic Equations**

Default Premium

δ	φ_1	φ_2	θ_1	θ_2	R^2	Abs. Roots
0.066	0.938		0.381		0.936	0.938
(0.681)	(66.110)		(9.845)			
0.066	0.942	−0.004		0.377	0.936	0.938, 0.004
(0.681)	(8.332)	(−0.037)		(3.566)		
0.067	0.938		0.381	0.002	0.936	0.938
(0.685)	(61.429)		(8.760)	(0.038)		

Real GDP Growth Rate

δ	φ_1	φ_2	φ_3	θ_1	θ_2	R^2	Abs. Roots
1.512	0.515			−0.174		0.137	0.515
(4.319)	(4.909)			(−1.531)			
2.851	−0.234	0.318		0.563		0.148	0.693, 0.459
(8.174)	(−0.839)	(3.418)		(1.950)			
2.286	0.266			0.057	0.184	0.151	0.266
(6.627)	(1.566)			(0.345)	(2.516)		
1.185	1.689	−1.292	0.224	−1.376	0.921	0.177	0.975, 0.975, 0.235
(4.101)	(21.579)	(−11.962)	(3.340)	(−29.644)	(16.714)		

The right panels of Figure 10.7 show a simulated path and the ACF and PACF from an ARMA(2,2). This path uses the same shocks as in the top-left panel, only with a different model. Both the ACF and the PACF have an oscillating pattern and decay toward zero as the lag length increases.

To summarize, an AR(p) model has a slowly decaying ACF and a PACF that abruptly truncates at lag p; an MA(q) model has an ACF that abruptly truncates at lag q and a PACF that slowly decays; an ARMA(p,q) has an ACF and PACF that both slowly decay.

10.7 SAMPLE AUTOCORRELATION

Sample autocorrelation and partial autocorrelations are used to build and validate ARMA models. These statistics are first applied to the data to understand the dependence structure and to select a set of candidate models. They are then applied to the estimated residuals $\{\hat{\epsilon}_t\}$ to determine whether they are consistent with the key assumption that $\epsilon_t \sim WN(0, \sigma^2)$.

The most common estimator of the sample autocovariance is

$$\hat{\gamma}_h = (T - h)^{-1} \sum_{i=h+1}^{T} (Y_i - \bar{Y})(Y_{i-h} - \bar{Y}), \qquad (10.31)$$

where $\bar{Y}$ is the full sample average. This estimator uses all available data to estimate $\hat{\gamma}_h$. Note that the first h data points are lost in creating the required h lags so that the index i begins at $h + 1$.

The autocorrelation estimator is then defined as:

$$\hat{\rho}_h = \frac{\sum_{i=h+1}^{T} (Y_i - \bar{Y})(Y_{i-h} - \bar{Y})}{\sum_{i=1}^{T} (Y_i - \bar{Y})^2} = \frac{\hat{\gamma}_h}{\hat{\gamma}_0} \qquad (10.32)$$

This estimator has a slight bias towards zero that disappears as T becomes large.

Joint Tests of Autocorrelations

Testing autocorrelation in the residuals is a standard specification check applied after fitting an ARMA model. It is common practice to use both graphical inspections as well as formal tests in specification analysis.

Graphical examination of a fitted model includes plotting the residuals to check for any apparent deficiencies and plotting the sample ACF and PACF of the residuals (i.e., $\hat{\epsilon}_t$). For example, a graphical examination can involve assessing the 95% confidence intervals for the null H_0: $\rho_h = 0$ (ACF) or H_1: $\alpha_h = 0$ (PACF). The presence of many violations of the 95% confidence interval in the sample ACF or PACF would indicate that the model does not adequately capture the dynamics of the data. However, it is common to see a few estimates outside of the 95% confidence bands, even in a well-specified model. This pattern makes it challenging to rely exclusively on graphical tools to determine whether a model is well-specified.

Two closely related joint tests of autocorrelations are often used when validating a model. They both test the joint null hypothesis that all of the autocorrelations are simultaneously zero (i.e., H_0: $\rho_1 = \rho_2 = \cdots = \rho_h = 0$) against the alternative hypothesis that at least one is non-zero (i.e., H_1: $\rho_j \neq 0$ for some j). Both tests have asymptotic χ_h^2 distributions. Values of the test statistic larger than the critical value indicate that the autocorrelations are not zero.

The Box-Pierce test statistic is the sum of the squared autocorrelations scaled by the sample size T:

$$Q_{BP} = T \sum_{i=1}^{h} \hat{\rho}_i^2 \qquad (10.33)$$

When the null is true, $\sqrt{T}\hat{\rho}_i$ is asymptotically distributed as a standard normal and thus $T\hat{\rho}_i^2$ is distributed as a χ_1^2. The estimated autocorrelations are also independent and so the sum of h independent χ_1^2 variables is, by definition, a χ_h^2 variable.

The Ljung-Box statistic is a modified version of the Box-Pierce statistic that works better in smaller samples. It is defined as:

$$Q_{LB} = T \sum_{i=1}^{h} \left(\frac{T + 2}{T - i} \right) \hat{\rho}_i^2 \qquad (10.34)$$

so that in large samples $Q_{BP} \approx Q_{LB}$. The Ljung-Box test is also distributed as a χ_h^2.

When T is modest (e.g., <100), the finite sample distribution of the Ljung-Box when the null is true is closer to the asymptotic χ_h^2 distribution. Therefore, it is the preferred method to test multiple autocorrelations.

The choice of h can affect the conclusion of the test that the residuals have no autocorrelation. When testing the specification of an ARMA(p,q) model, h should be larger than max(p,q). Residual autocorrelations that fall within the lags of the model are typically small, and the ARMA(p,q) estimator is effectively overfitting these values. It is common to use between 5 and 20 lags when testing a small ARMA model ($p, q \leq 2$).

Specification Analysis

Autocorrelation and partial autocorrelation play a key role in both model selection and specification analysis. The left column of Figure 10.8 plots 24 sample autocorrelations and partial autocorrelations of the default premium (i.e., two years of monthly observations). The right column shows 12 sample ACF and PACF values of real GDP growth (i.e., three years of quarterly observations). The horizontal dashed lines mark the boundary of

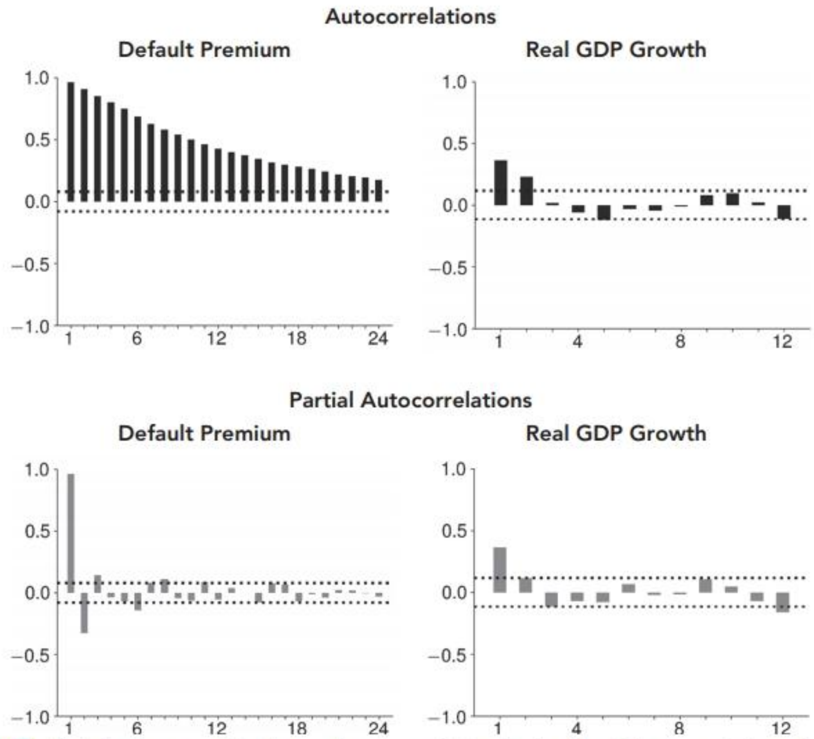


Figure 10.8 The left column contains the sample autocorrelations (top) and partial autocorrelations of the default premium. The right panel shows the sample autocorrelations and partial autocorrelations of real GDP growth.

a 95% confidence interval derived from an assumption that the process is white noise.

The ACF of the default premium decays very slowly towards zero, which is consistent with an AR model. The PACF has three values which are statistically different from zero. This pattern is consistent with either a low-order AR or an ARMA with a MA component that has a small coefficient.

The ACF and PACF of real GDP growth indicate weaker persistence, and the ACF is different from zero in the first two lags. The PACF resembles the ACF, and the first two lags are statistically different from zero. These are consistent with a low-order AR, MA, or ARMA.

Sample autocorrelations and partial autocorrelations are also applied to investigate the adequacy of model specifications. In Figure 10.9, an ARMA(1,1) is estimated on the default premium data, and an AR(1) is used to model real GDP growth.

The top four panels in Figure 10.9 show the sample ACF and PACF of the estimated model residuals from these specifications. Note that these plots are the equivalents of those in Figure 10.8, except that the data are replaced by the model residuals. All four plots indicate that the residuals are consistent with a white noise process. The estimated autocorrelations are generally insignificant, small in magnitude, and lack any discernible pattern.

The bottom panel plots both the Ljung-Box statistic and its p-value. The statistic is computed for all lag lengths between 1 and 12. The default premium residuals have small autocorrelations in the first four lags, and the null that all autocorrelations

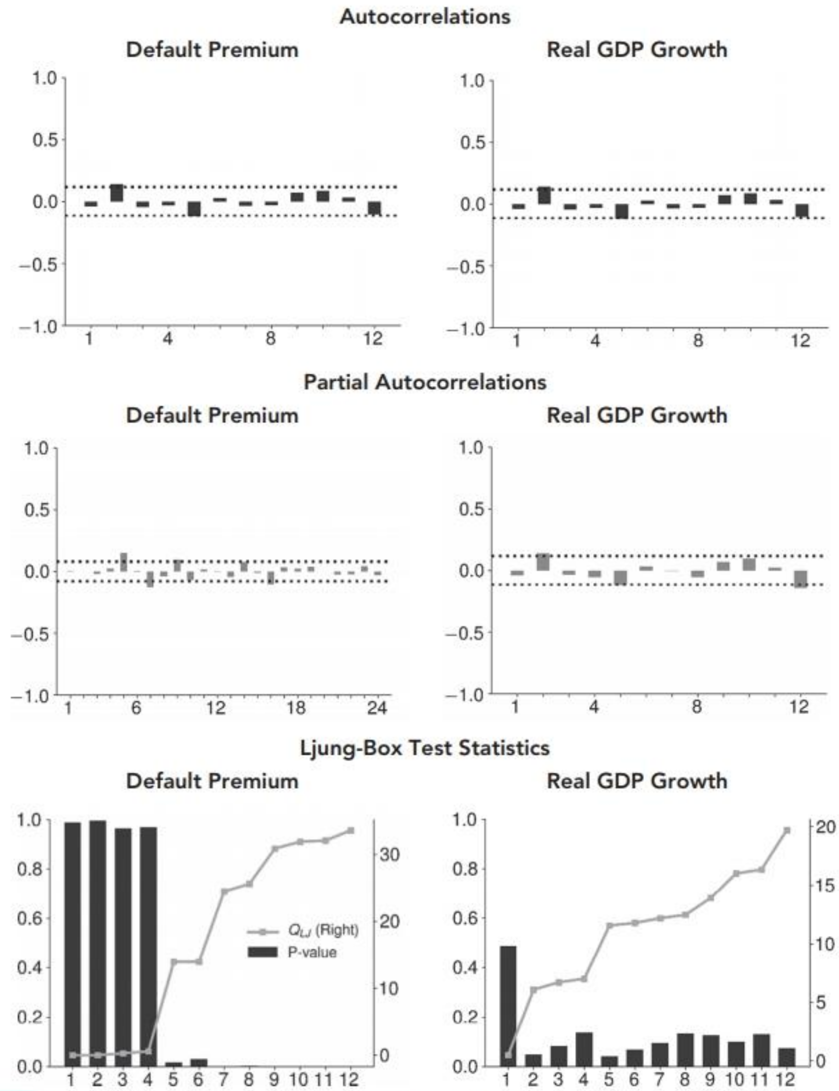


Figure 10.9 The top panels show the sample ACF and PACF of the model residuals. The residuals for the default premium are from an ARMA(1,1). The residuals for real GDP growth are from an AR(1). The bottom plot shows the Ljung-Box Q statistic (line, right axis) and its p-value (bars, left-axis).

PARAMETER ESTIMATION IN AR AND ARMA MODELS

The parameters of an AR(p) model can be estimated using OLS. The dependent variable is Y_t and the explanatory variables are the p lagged values, which are all observable. The parameter estimators are defined as:

$$\underset{\delta,\phi_1,\ldots\phi_p}{\arg\min} \sum_{t=p+1}^{T} (Y_t - \delta - \phi_1 Y_{t-1} - \cdots - \phi_p Y_{t-p})^2 \quad (10.35)$$

Estimation of the parameters in an AR is identical to parameter estimation in linear regression models. Standard errors are also estimated using the same expressions as in a linear regression. The only feature of a linear regression that cannot be directly applied to parameter estimation in an AR(p) is unbiasedness. Whereas the linear regression parameter estimators are unbiased, parameter estimators of the AR coefficients are biased (although they are consistent).

This bias comes from the fact that Y_t's innovation (i.e., ϵ_t) can appear on both sides of the equation: on the left when Y_t is the dependent variable and on the right when Y_t is a lagged value for $Y_{t+1}, Y_{t+2}, \ldots, Y_{t+p}$. This repetition of the same shock on both sides is the source of the finite-sample bias.

OLS cannot be used when an MA component is added to an AR or when estimating a pure MA. Whereas OLS requires that all explanatory variables are observable, the lagged errors that appear on the right-hand side of an MA do not meet this requirement and can only be inferred once the model parameters have been estimated. The coefficients of MA and ARMA models are instead estimated using maximum likelihood, which assumes that the innovations are jointly normally distributed with mean zero and variance σ^2.

are zero is not rejected. However, the 5th lag autocorrelation of the estimated residuals is large enough that the Q-statistic rejects the null. The remainder of the lags also indicate that the null should be rejected.[7]

The residuals from the AR(1) estimated on the real GDP growth data are consistent with white noise. Both the sample ACF and PACF have one value that is statistically significant. However, a small number of rejections are expected when examining multiple lags because each has a 5% chance of rejecting when the null is true. In other words, when testing the autocorrelations at many lags individually, some will be statistically significant purely by chance alone. The Ljung-Box test also does not reject the null at any lag length.

10.8 MODEL SELECTION

Determining the appropriate lag lengths for the AR and MA components (i.e., p and q, respectively) is a key challenge when building an ARMA model. The first step in model building is to inspect the sample autocorrelation and sample PACFs. These provide essential insights into the correlation structure of the data and can be used to determine the class of models that are likely to explain that structure.

For example, the ACF and PACF of an AR(p) model have a distinct pattern regression: It decays slowly (possibly with oscillations), while the PACF has a sharp cutoff at p lags. If the sample counterparts of these exhibit a similar pattern, then an AR model is likely to fit the data better than an MA or ARMA.

In general, slow decay in the sample ACF indicates that the model needs an AR component, and slow decay in the sample PACF indicates that the model should have an MA component. Note that the goal of this initial step is to identify plausible candidate models. It is *not* necessary to make a final selection using only graphical analysis.

Once an initial set of models has been identified, the next step is to measure their fit. The most natural measure of fit is the sample variance of the estimated residuals, also known as the Mean Squared Error of the model. It is defined as:

$$\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}\hat{\epsilon}_t^2 \quad (10.36)$$

Smaller values indicate that the model explains more of the time series. Unfortunately, choosing a model to minimize the residual variance $\hat{\sigma}^2$ also selects a specification that is far too large. This objective suffers from the same problem as maximizing the R^2 in a linear regression—adding additional lags to either the AR or MA components always produces smaller MSEs. The solution to this overfitting problem is to add a penalty to the MSE that increases each time a new parameter is added.

Penalized MSE measures are known as information criteria (IC). The leading two among these are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC).[8]

[7] The Ljung-Box test is not robust with respect to heteroskedasticity in the data. A robust version can be constructed by estimating an AR(p) on the residuals using OLS and using White's heteroskedasticity robust estimator. The regression is:

$$\hat{\epsilon}_t = \delta + \phi_1\hat{\epsilon}_{t-1} + \phi_2\hat{\epsilon}_{t-2} + \cdots + \phi_p\hat{\epsilon}_{t-p} + \eta_t$$

The null hypothesis is that all coefficients on the lagged terms are zero, H_0: $\phi_1 = \cdots = \phi_p = 0$, and that alternative is H_1: $\phi_j \neq 0$ for some j.

[8] This is also known as the Schwarz IC (SIC or SBIC).

These IC both reflect the bias-variance tradeoff and attempt to balance the cost and benefit of alternative specifications that include different orders of the AR and MA components.

The AIC is defined as:

$$AIC = T \ln \hat{\sigma}^2 + 2k, \quad (10.37)$$

where T is the sample size and k is the number of parameters.

The BIC alters the penalty and is defined as:

$$BIC = T \ln \hat{\sigma}^2 + k \ln T \quad (10.38)$$

Note that the penalty term in the AIC adds a constant "cost" of two per parameter, whereas the penalty in the BIC has a cost that *slowly* increases with T. This difference in the penalty has two implications.

1. The BIC always selects a model that is no larger (i.e., in terms of the lag lengths p and q) than the model selected by the AIC (assuming $T \geq 8$, because $\ln 8 = 2.07$).
2. The BIC is a consistent[9] model selection criterion (i.e., the true model is selected as $T \rightarrow \infty$).

The difference between the two IC can be understood in the context of hypothesis testing.

- The AIC behaves like a model selection methodology that includes any variable that is statistically significant with a fixed-test size of s%. When using a fixed-test size, there is an s% chance that a coefficient that is not relevant is selected. In aggregate, this can lead to selecting models that are too large.
- The BIC behaves in the same way, only $s \rightarrow 0$ as $T \rightarrow \infty$. The quantity s goes to zero slowly, so that any relevant variable always has a t-statistic that is larger than the required critical value in large samples. Variables that are not needed, however, always have t-statistics that are less than the critical value (for a T large enough). As a result, such variables are always excluded.

Box-Jenkins

Box and Jenkins (1976)[10] proposed a formal model selection approach for specifying and estimating the most appropriate model from the ARMA family. Their framework comprised three steps:

1. Identifying the appropriate model using ACF and PACF plots
2. Estimating the model parameters using OLS (for AR only) or maximum likelihood (for AR, MA or ARMA)
3. Diagnostic checking of the proposed model's residuals.

It is possible to specify two (or more) models that have different parameter values but are equivalent. Equivalence between ARMA models means that the mean, ACF, and PACF of the models are identical.

The Box-Jenkins methodology provides two principles to select among equivalent models.

1. **Parsimony**—When selecting between models that are equivalent (and thus produce the same fit), always choose the model with fewer parameters.
2. **Invertibility**—When choosing parameters in MA processes (either pure MA or ARMA), always select parameter values so that the MA coefficients are invertible. Because any MA(q) model has 2^q equivalent representations, requiring invertibility is a simple method to choose one of these representations.

These two principles are enough to ensure that a unique specification is selected to match any ACF and PACF.

10.9 FORECASTING

Forecasts use current information to predict the future. While it is common to make one-step-ahead forecasts, which predict the next value in the time series, forecasts can be generated for any horizon h.

The one-step forecast $E[Y_{T+1}|\mathcal{F}_T]$ is the expectation of Y_{T+1} conditional on F_T, which is the time T information set. This information set contains all values that are known at time T, including the entire history of Y: $Y_T, Y_{T-1}, \ldots$. It also includes the history of the shocks $\epsilon_T, \epsilon_{T-1}, \ldots$ (even though these are not directly observed)[11] as well as all values of any other variable that occurred at time T or earlier. The shorthand notation $E_T[\cdot]$ is commonly used for the conditional expectation where:

$$E_T[Y_{T+1}] = E[Y_{T+1}|\mathcal{F}_T]$$

Three rules simplify recursively generating forecasts.

1. The expectation of any variable with a time subscript T or earlier is the realization of that variable (e.g., $E_T[Y_T] = Y_T$). Residuals are also in the information set (e.g., $E_T[\epsilon_{T-2}] = \epsilon_{T-2}$).
2. The expectation of future shocks is zero (e.g., $E_T[\epsilon_{T+1}] = 0$). This is true for any horizon (i.e., $E_T[\epsilon_{T+h}] = 0$ for all $h \geq 1$).

[9] This is analogous to parameter estimator consistency, which ensures that estimated parameters converge to their true values.

[10] Box, G.E.P. and Jenkins, G.M. (1976) *Time Series Analysis: Forecasting and Control*, 2nd Edition, Holden Day, San Francisco.

[11] While the shocks are not directly observed, they can be estimated from the data by transforming the ARMA process into the AR(∞) $y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \cdots = \epsilon_t$. The appendix contains an example illustrating how an MA(1) process is inverted into an AR(∞).

3. Forecasts are generated recursively starting with $E_T[Y_{T+1}]$. The forecast at horizon h may depend on forecasts from earlier steps ($E_T[Y_{T+1}], \ldots, E_T[Y_{T+h-1}]$). When these quantities appear in the forecast for period $T + h$, they are replaced by the forecasts computed for horizons $1, 2, \ldots, h - 1$.

APPLIED MODEL SELECTION

Figure 10.10 plots the AIC and the BIC values for ARMA models where $p \leq 4$ and $q \leq 2$ for both the default premium and real GDP data. The y-axis labels have been omitted because only the relative values matter for an information criterion.

The AIC and the BIC both select the same model for the default premium—an ARMA(1,1). The two criteria also agree on the second and third best specifications, which are an ARMA(1,2) and an ARMA(2,1), respectively. In other words, the lack of agreement between the two criteria for real GDP growth reflects the earlier observation that none of the models fit particularly well in this case.

The IC disagree on the preferred model for the real GDP growth data. The AIC selects an ARMA(3,2), whereas BIC prefers an ARMA(1,0) [i.e., simply an AR(1)]. The second and third best models according to the AIC are an ARMA(2,2) and an AR(3), respectively. The BIC ranks an MA(2) as the second-best model and an AR(2) as the third. The differences between the criteria reflect the effect of the sample size and the precision of the estimates in models for real GDP growth.

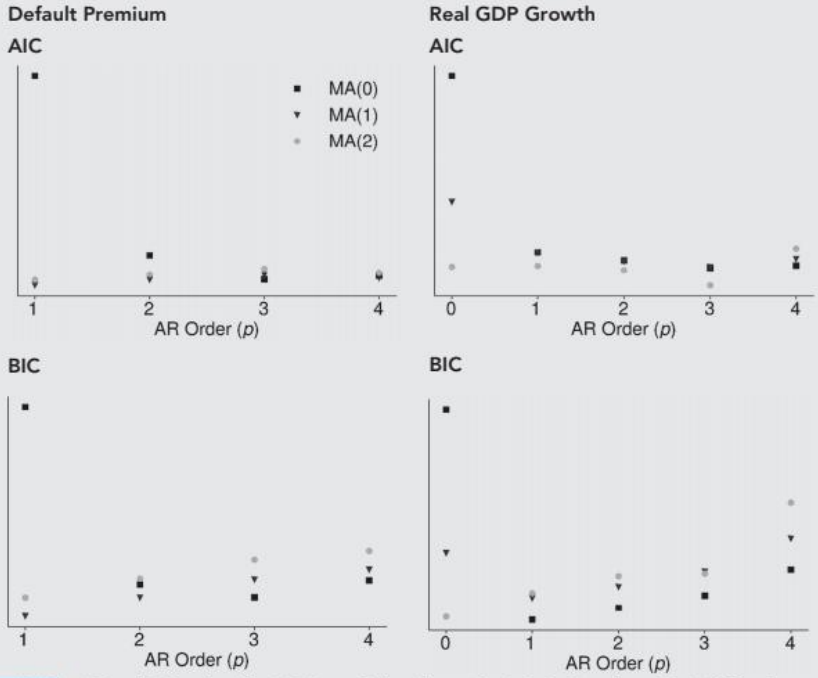


Figure 10.10 AIC and BIC values for ARMA models with $p \in \{0, 1, 2, 3, 4\}$ and $q \in \{0, 1, 2\}$ for the default premium (left) and real GDP growth (right). Models with $p = 0$ have been excluded from the default premium plot because these produce excessively large values of the ICs. The selected model has the smallest information criterion.

Applying these rules to an AR(1), the one-step forecast is:

$$\begin{aligned} E_T[Y_{T+1}] &= E_T[\delta + \phi Y_T + \epsilon_{T+1}] \\ &= \delta + \phi E_T[Y_T] + E_T[\epsilon_{T+1}] \\ &= \delta + \phi Y_T \end{aligned}$$

and depends only on the final observed value Y_T.

The two-step forecast is:

$$\begin{aligned} E_T[Y_{T+2}] &= E_T[\delta + \phi Y_{T+1} + \epsilon_{T+2}] \\ &= \delta + \phi E_T[Y_{T+1}] + E_T[\epsilon_{T+2}] \\ &= \delta + \phi(\delta + \phi Y_T) \\ &= \delta + \phi\delta + \phi^2 Y_T \end{aligned}$$

The two-step forecast depends on the one-step forecast and so is also a function of Y_T.

These steps repeat for any horizon h so that:

$$\begin{aligned} E_T[Y_{T+h}] &= \delta + \phi\delta + \phi^2\delta + \cdots + \phi^{h-1}\delta + \phi^h Y_T \\ &= \sum_{i=0}^{h} \phi^i \delta + \phi^h Y_T \end{aligned} \qquad (10.39)$$

When h is large, ϕ^h must be small because $\{Y_t\}$ is assumed to be a stationary time series. It can be shown that:

$$\lim_{h \to \infty} \sum_{i=0}^{h} \phi^i \delta + \phi^h Y_T = \frac{\delta}{1 - \phi}$$

This limit is the same as the mean reversion level (or long-run mean) of an AR(1).

This duality between the mean reversion level and the long run forecast reflects a property of any covariance-stationary time series—the current value of Y_T always has a negligible impact on the values of Y in the distant future. Formally, for any covariance-stationary time series:

$$\lim_{h \to \infty} E_T[Y_{T+h}] = E[Y_t],$$

which is the long-run (or unconditional) mean. So in other words, the forecasts from any stationary AR(p) model converge upon the unconditional mean of the series as the forecast horizon h increases.

These same steps generalize to MA and ARMA processes using the three rules of forecasting. For example, the first three forecasts from an MA(2):

$$Y_T = \mu + \theta_1 \epsilon_{T-1} + \theta_2 \epsilon_{T-2} + \epsilon_T$$

are:

$$\begin{aligned} E[Y_{T+1}] &= \mu + \theta_1 \epsilon_T + \theta_2 \epsilon_{T-1} \\ E[Y_{T+2}] &= \mu + \theta_2 \epsilon_T \\ E[Y_{T+3}] &= \mu \end{aligned}$$

MAs depend on (at most) q lags of the residual, and so all forecasts for horizon $h > q$ are the long-run mean μ.

The forecast error is the difference between the realized (future) value and its time T forecast and is defined as $Y_{T+h} - E_T[Y_{T+h}]$. The one-step forecast error:

$$Y_{T+1} - E_T[Y_{T+1}] = \epsilon_{T+1}$$

is always just the surprise in the next period in any ARMA model. Forecasts errors for longer-horizon forecasts are functions of the model parameters.

APPLIED FORECASTING

The preferred model for the default premium is the ARMA(1,1):

$$DEF_t = 0.066 + 0.938\, DEF_{t-1} + 0.381\, \epsilon_{t-1} + \epsilon_t$$

Generating out-of-sample forecasts requires the final value of the default premium and the final estimated residual. These values are $DEF_T = 1.11$ and $\hat{\epsilon}_T = 0.0843$. The one-step ahead forecast is:

$$E_T[DEF_{T+1}] = 0.066 + 0.938 \times 1.11 + 0.381 \times 0.0843 = 1.139$$

The two-step and higher forecasts are recursively computed using the relationship:

$$E_T[DEF_{T+h}] = 0.066 + 0.938\, E_T[DEF_{T+h-1}]$$

These values are reported in Table 10.4.The AR parameter is relatively close to 1, and so the forecasts converge slowly to the long-run mean of 0.066/(1 − .938) = 1.06.

The selected model for real GDP growth is the AR(1):

$$RGDPG_t = 1.996 + 0.360\, RGDPG_{t-1} + \epsilon_t.$$

The final real GDP growth rate is $RGDPG_T = 2.564$ in 2018Q4. Forecasts for longer horizons are recursively computed using the relationship:

$$E_T[RGPDG_{T+h}] = 1.996 + 0.360\, E_T[RGDPG_{T+h-1}]$$

For example, the first forecast is 1.996 + 0.36 × 2.564 = 2.919, and the second is 1.996 + 0.360 × 2.919 = 3.047. The remaining forecasts for the first year are presented in Table 10.4. The persistence parameter in the AR(1) is small, and the forecasts rapidly converge to the long-run mean of the growth rate of real GDP (i.e., 1.996/(1 − 0.36) = 3.119).

Table 10.4 Multi-Step Forecasts for the Default Premium (Left) and Real GDP Growth Rate (Right). The Forecasts for the Default Premium Are Constructed from an ARMA(1,1). The Forecasts for the Real GDP Growth Rate Are from an AR(1) Model

	Default Premium			Real GDP Growth
Horizon	Date	Value	Date	Value
1	Jan 2019	1.139	2019Q1	2.919
2	Feb 2019	1.135	2019Q2	3.047
3	Mar 2019	1.131	2019Q3	3.093
4	Apr 2019	1.127	2019Q4	3.110
5	May 2019	1.123	2020Q1	3.116
6	Jun 2019	1.119	2020Q2	3.118
7	Jul 2019	1.116	2020Q3	3.119
8	Aug 2019	1.113	2020Q4	3.119

10.10 SEASONALITY

Macrofinancial time series often have seasonalities. For example, many activities related to home buying and construction are higher in the summer months than in the winter months. Seasonality can be deterministic (e.g., constant level shift in a series during the summer months) or stochastic. Series with deterministic seasonality are non-stationary, whereas those with stochastic seasonality *can be* stationary (i.e., they can be modeled using ARMAs).

Macrofinancial seasonality occurs on an annual basis. For example, the seasonal component for quarterly data appears in gaps of four and that of monthly data appears in gaps of 12.

A pure seasonal model only uses lags at the seasonal frequency. For example, a pure seasonal AR(1) model of quarterly data series is:

$$(1 - \phi L^4)Y_t = \delta + \epsilon_t$$
$$Y_t = \delta + \phi Y_{t-4} + \epsilon_t$$

This process has peculiar dynamics that are identical to those generated from four independent AR(1) processes interwoven to make an AR(4). However, this structure is not plausible in most economic time series.

A more plausible structure includes both short-term and seasonal components. The seasonal component uses lags at the seasonal frequency, while the short-term component uses lags at the observation frequency. A seasonal ARMA combines these two components into a single specification. For example, a model using monthly data with a seasonal AR(1) and a short-term AR(1) is:

$$(1 - \phi_1 L)(1 - \phi_s L^{12})Y_t = \delta + \epsilon_t$$
$$(1 - \phi_1 L - \phi_s L^{12} + \phi_1\phi_s L^{13})Y_t = \delta + \epsilon_t$$
$$Y_t = \delta + \phi_1 Y_{t-1} + \phi_s Y_{t-12} + \phi_1\phi_s Y_{t-13} + \epsilon_t$$

This seasonal AR model is a restricted AR(13), where most of the coefficients are zero, and the coefficient on lag 13 is linked to the coefficient on the other included lags.

Seasonality can be introduced to the AR component, the MA component, or both. The seasonality is included in a model by multiplying the short-run lag polynomial by a seasonal lag polynomial. The specification of Seasonal ARMA models is denoted:

$$\text{ARMA}\,(p, q) \times (p_s, q_s)_f$$

where p and q are the orders of the short-run lag polynomials, p_s and q_s are the orders of the seasonal lag polynomials, and f is the seasonal horizon (e.g., 4 or 12). For example, the seasonal AR in the previous example is an ARMA(1,0) $\times$ $(1,0)_{12}$.

In practice, seasonal components are usually restricted to one lag because the precision of the parameters related to the seasonal component depends on the number of full seasonal cycles in the sample. In turn, the number of seasonal observations depends on the sample size and the frequency (i.e., T/f). For example, when T is modest (e.g., 100), then T/f is small.

MODEL SELECTION IN SEASONAL TIME SERIES

Model selection in seasonal time series is identical to selection in non-seasonal time series. The ACF and PACF are inspected at both the observational and the seasonal frequencies to determine an initial model specification.

Seasonal ARs have slowly decaying ACFs and a sharp cutoff in the PACF (when using the seasonal frequency). Seasonal MAs have the opposite pattern, where the PACF slowly decays and the ACF drops off sharply. Dynamics present in the ACF and PACF at the observation frequency indicate that a short-term component is required. Finally, IC can be used to select between models that contain only observation frequency lags [i.e., an ARMA(p, q) $\times$ $(0,0)_f$)], only seasonal lags [i.e., ARMA(0,0) $\times$ $(p_s, q_s)_f$], or a mixture of both.

10.11 SUMMARY

This chapter introduces the tools used to analyze, model, and forecast time series. The focus has been exclusively on covariance-stationary time series that have constant means and autocovariances. This property ensures that the dependence is stable over time and is essential when modeling a time series.

Stationary series are modeled using autoregressive (AR) processes or moving averages (MA). These models imply different dependence structures, but both are built using white noise shocks that are uncorrelated over time. This assumption allows model residuals to be inspected to determine whether a specification appears to be adequate. In addition, these models are often combined in autoregressive moving average models (ARMA), which can be used to approximate any linear process.

Alternative specifications are compared using one of two IC: the AIC or the BIC. These criteria tradeoff bias and variance by penalizing larger models that do not improve fit. This BIC leads to consistent model selection and is more conservative than the AIC.

Finally, this chapter explained how ARMA models are used to produce out-of-sample forecasts. These concepts extend directly to time series with seasonal components using seasonal ARMA models, which build models by combining observation and seasonal frequencies.

APPENDIX

Characteristic Equations

In higher order polynomials, it is difficult to define the values of the model coefficients where the polynomial is invertible—and hence the parameter values where an AR is covariance-stationary. The stationarity conditions can be determined by associating the p^{th} order lag polynomial:

$$1 - \phi_1 L - \cdots - \phi_p L^P \quad (10.40)$$

with its characteristic equation:

$$1 - \phi_1 z - \ldots - \phi_p z^P = 0 \quad (10.41)$$

The characteristic equation is also a p^{th} order polynomial where the ordering of the powers has been reversed.

The lag polynomial is invertible if the roots of the characteristic equation $c_1, c_2 \ldots, c_p$ in the expression:

$$(z - c_1)(z - c_2)\ldots(z - c_p) = 0 \quad (10.42)$$

are all less than 1 in absolute value.[12]

In the first-order polynomial:

$$1 - \phi_1 L$$

the associated characteristic equation is

$$1 - \phi_1 z = 0$$

and the solution is $z = 1/\phi_1$. If $|\phi_1| > 1$, then the polynomial is invertible and the AR(1) is stationary.

In the AR(2):

$$Y_t = \delta + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \epsilon_t$$

The lag polynomial is $1 - \phi_1 L - \phi_2 L^2$ and its associated characteristic equation is $1 - \phi_1 z - \phi_2 z^2 = 0$. If $\phi_1 = 1.4$ and $\phi_2 = -0.45$, then the factored characteristic polynomial is $(0.9z - 1)(0.5z - 1)$. The two roots are 1/0.9 and 1/0.5. Both are greater than 1 in absolute value, and so this AR(2) is stationary.

[12] The roots may be complex valued, so that $c_j = a_j + b_j\sqrt{-1}$. When c_j is complex, $|c_j| = \sqrt{a_j^2 + b_j^2}$ is known as the complex modulus.

The following questions are intended to help candidates understand the material. They are *not* actual FRM exam questions.

QUESTIONS

Short Concept Questions

10.1 What are the three components of a time series?

10.2 How does Gaussian white noise differ from general white noise?

10.3 What are the properties of the lag operator?

10.4 What are the key features of the ACF and PACF of an MA(q) process?

10.5 What is the maximum number of non-zero autocorrelations in an MA(q)?

10.6 Rewrite the MA(2), $Y_t = 0.3 + 0.8\epsilon_{t-1} + 0.16\epsilon_{t-2} + \epsilon_t$, using a lag polynomial.

10.7 Why does the AIC select a model that is either the same size or larger than the model selected using the BIC ?

10.8 What steps should be taken if the ACF and/or PACF of model residuals do not appear to be white noise?

Practice Questions

10.9 In the covariance-stationary AR(2), $Y_t = 0.3 + 1.4Y_{t-1} - 0.6Y_{t-2} + \epsilon_t$, where $\epsilon_t \sim WN(0, 0.3^2)$, what is the long-run mean $E[Y_t]$ and variance $V[Y_t]$?

10.10 If $(1 - 0.4L)(1 - 0.9L^4)Y_t = \epsilon_t$, what is the form of this model if written as a standard AR(5) with $Y_{t-1}, \ldots, Y_{t-5}$ on the right-hand side of the equation?

10.11 For the equation:

$$Y_t = \frac{15}{32} + \frac{5}{4}Y_{t-1} - \frac{3}{8}Y_{t-2} + \epsilon_t$$

What is the lag polynomial?

10.12 Given the following data for a set of innovations:

	ϵ_t
1	−0.58
2	1.62
3	−1.34
4	0.88
5	−1.11
6	0.52
7	0.81
8	−0.65
9	0.85
10	−0.88

a. Calculate the time series (i.e., calculate y_1, y_2, etc.) for the following AR(1) model, taking $y_0 = 0$, $\delta = 0.5$, and $\phi = 0.75$:

b. Calculate the time series for the following MA(1) model, taking $\epsilon_0 = 0$, $\mu = 0.5$, and $\theta = 0.75$:

c. Calculate the time series for the following ARMA(1,1) model, taking $\epsilon_0 = 0$, $q_0 = 0$, $\eta = 0.5$, and $\phi = \theta = 0.375$:

10.13 For the MA(2), $Y_t = 4.1 + 5\epsilon_{t-1} + 6.25\epsilon_{t-2} + \epsilon_t$, where $\epsilon_t \sim WN(0, \sigma^2)$, what is the ACF?

10.14 Suppose that all residual autocorrelations are $1.5/\sqrt{T}$, where T is the sample size. Would these violate the confidence bands in an ACF plot?

10.15 Suppose that you observed sample autocorrelations $\hat{\rho}_1 = 0.24$, $\hat{\rho}_2 = -0.04$, and $\hat{\rho}_3 = 0.08$ in a time series with 100 observations. Would a Ljung-Box Q statistic reject its null hypothesis using a test with a size of 5%? Would the test reject the null using a size of 10% or 1%?

10.16 Suppose that you fit a range of ARMA models to a data set. The ARMA orders and estimated residual variance $\hat{\sigma}^2$ are

Order	σ^2	Order	σ^2	Order	σ^2
(1, 0)	1.212	(1, 1)	1.147	(1, 2)	1.112
(2, 0)	1.175	(2, 1)	1.097	(2, 2)	1.097
(3, 0)	1.150	(3, 1)	1.097	(3, 2)	1.068
(4, 0)	1.133	(4, 1)	1.096	(4, 2)	1.065

Which model is selected by the AIC and BIC if the sample size $T = 250$?

10.17 The 2018Q4 value of real GDP growth is $RGDPG_{T-1} = 2.793$. What are the forecasts for 2019Q1 – 2019Q4 using the AR(2) model in Table 10.1?

The following questions are intended to help candidates understand the material. They are ***not*** actual FRM exam questions.

ANSWERS

Short Concept Questions

10.1 The three components are trend, seasonal, and cyclical.

10.2 General white noise can have any distribution and does not have the be independent over time. Gaussian white noise is iid, and each shock is normally distributed $N(0, \sigma^2)$.

10.3 **a.** The lag operator shifts the time index back one observation.

b. Lag operators multiply so that $LL = L^2$ and $L^2Y_t = L(LY_t) = LY_{t-1}$.

c. The lag operator applied to a constant is just the constant, $L\delta = \delta$.

d. The pth order lag polynomial is written $\phi(L) = 1 + \phi_1 L + \phi_2 L^2 + \cdots + \phi_p L^P$.

e. Lag polynomials can be multiplied to produce another lag polynomial.

f. Under some assumptions, lag polynomials can be inverted.

10.4 The ACF of an MA(q) is non-zero until q lags, and then is zero for all values above q. The PACF declines slowly in the lag and may oscillate, but there is no lag above which the PACF is always zero.

10.5 q, the order of the model.

10.6 $Y_t = 0.3 + \epsilon_t(1 + 0.8L + 0.16L^2)$

10.7 The BIC has a sharper penalty for all sample sizes larger than 7, and so will always prefer fewer lags.

10.8 The model should be expanded to include additional AR or MA coefficients.

Solved Problems

10.9 Because this process is covariance-stationary

$$E[Y_t] = \mu = \frac{0.3}{1 - 1.4 - (0.6)} = 1.5$$

$$V[Y_t] = \gamma_0 = \frac{0.3^2}{1 - 1.4 - (0.6)} = 0.45$$

10.10 $(1 - 0.4L)(1 - 0.9L^4)Y_t = (1 - 0.4L - 0.9L^4 + 0.36L^5)Y_t$

$$\Rightarrow Y_t - 0.4Y_{t-1} - 0.9Y_{t-4} + 0.36Y_{t-5} = \epsilon_t$$

$$Y_t = 0.4Y_{t-1} + 0.9Y_{t-4} - 0.36Y_{t-5} + \epsilon_t$$

10.11 $1 - \phi_1 L - \phi_2 L^2 = 1 - \frac{5}{4}L + \frac{3}{8}L^2$

10.12 **a.** $Y_t = \delta + \phi_{Y_{t-1}+\epsilon_t}$

	y_t
1	−0.08
2	2.06
3	0.71
4	1.91
5	0.82
6	1.64
7	2.54
8	1.76
9	2.67
10	1.62

Each value is generated via the formula:

$$y_1 = 0.5 + 0.75 * 0 - 0.58 = -0.08$$

$$y_2 = 0.5 + 0.75 * (-0.08) + 1.62 = 2.12$$

and so forth.

b. $z_t = \mu + \theta\epsilon_{t-1} + \epsilon_t$

	z_t
1	−0.08
2	1.69
3	0.38
4	0.38
5	0.05
6	0.19
7	1.70
8	0.46
9	0.86
10	0.26

Each value is generated via the formula:

$$z_1 = 0.5 + 0.75 * 0 - 0.58 = -0.08$$

$$z_2 = 0.5 + 0.75(-0.58) + 1.62 = 1.69$$

and so forth.

c. $q_t = n + \phi q_{t-1} + \theta\epsilon_{t-1} + \epsilon_t$

	q_t
1	−0.08
2	1.87
3	0.47
4	1.05
5	0.12
6	0.65
7	1.75
8	0.81
9	0.41
10	0.47

Each value is generated via the formula:

$$q_1 = 0.5 + 0.375 * 0 + 0.375 * 0 - 0.58 = -0.08$$

$$z_2 = 0.5 + 0.375(-0.08) + 0.375(-0.58) + 1.62 = 1.87$$

and so forth.

10.13 $\gamma_0 = (1 + 5^2 + 6.25^2)\sigma^2$, $\gamma_1 = (5 + 5 \times 6.25)\sigma^2$ and $\gamma_2 = 6\sigma^2$. The autocorrelations are then $\rho_0 = 1$, $\rho_1 = \gamma_1/\gamma_0 = 0.557$ and $\rho_2 = \gamma_2/\gamma_0 = 0.096$; $\rho_j = 0$ for $j > 2$ since this is an MA(2). Notice that σ^2 cancels out and so it is not necessary to know it.

10.14 No, the confidence bands are $\pm 1.96/\sqrt{T}$ and so these would not. If many autocorrelations are jointly relatively large, then this series is likely autocorrelated. A Ljung-Box test could be used to detect the joint autocorrelation.

10.15 The test statistic is

$$Q_{LB} = T\sum_{i=1}^{h}\left(\frac{T+2}{T-i}\right)\hat{\rho}_i^2$$

$$= 100\left(\frac{102}{99}\right)(0.24)^2 + 100\left(\frac{102}{98}\right)(-0.04)^2 + 100\left(\frac{102}{97}\right)(0.08)^2$$

$$= 5.93 + 0.16 + 0.67 = 6.77$$

The critical values for tests sizes of 10%, 5%, and 1% are 6.25, 7.81, and 11.34, respectively (CHISQ.INV(1-α, 5) in Excel, where α is the test size). The null is only rejected using a size of 10%.

10.16 The ICs and the selected model in **bold**:

Order	AIC	Order	AIC	Order	AIC
(1, 0)	50.068	(1, 1)	38.287	(1, 2)	32.540
(2, 0)	44.317	(2, 1)	29.145	(2, 2)	31.145
(3, 0)	40.940	(3, 1)	31.145	(3, 2)	**26.447**
(4, 0)	39.217	(4, 1)	32.917	(4, 2)	27.744

Order	BIC	Order	BIC	Order	BIC
(1, 0)	53.589	(1, 1)	45.330	(1, 2)	43.104
(2, 0)	51.360	(2, 1)	**39.709**	(2, 2)	45.231
(3, 0)	51.505	(3, 1)	45.231	(3, 2)	44.054
(4, 0)	53.303	(4, 1)	50.524	(4, 2)	48.872

10.17 All forecasts are recursively computed starting with the first:

$1.765 + 0.319 \times 2.564 + 0.114 \times 2.793 = 2.90.$

The two-step forecast uses the one-step forecast:

$1.765 + 0.319 \times 2.90 + 0.114 \times 2.564 = 2.98.$

The three- and four-step depend entirely on other forecasts:

$1.765 + 0.319 \times 2.98 + 0.114 \times 2.90 = 3.05.$

$1.765 + 0.319 \times 3.05 + 0.114 \times 2.98 = 3.08.$

11

Non-Stationary Time Series

Learning Objectives

After completing this reading you should be able to:

- Describe linear and nonlinear time trends.
- Explain how to use regression analysis to model seasonality.
- Describe a random walk and a unit root.
- Explain the challenges of modeling time series containing unit roots.
- Describe how to test if a time series contains a unit root.
- Explain how to construct an h-step-ahead point forecast for a time series with seasonality.
- Calculate the estimated trend value and form an interval forecast for a time series.

Covariance-stationary time series have means, variances, and autocovariances that do not depend on time. Any time series that is not covariance-stationary is non-stationary. This chapter covers the three most pervasive sources of non-stationarity in financial and economic time series: time trends, seasonalities, and unit roots (more commonly known as random walks).

Time trends are the simplest deviation from stationarity. Time trend models capture the propensity of many time series to grow over time. These models are often applied to the log transformation so that the trend captures the growth rate of the variable. Estimation and interpretation of parameters in trend models that contain no other dynamics are simple.

Seasonalities induce non-stationary behavior in time series by relating the mean of the process to the month or quarter of the year.[1] Seasonalities can be modeled in one of two ways: shifts in the mean that depend on the period of the year, or an annual cycle where the value in the current period depends on the shock in the same period in the previous year. These two types of seasonality are not mutually exclusive, and there are two approaches to modeling them. The simple method includes dummy variables for the month or quarter. These variables allow the mean to vary across the calendar year and accommodate predictable variation in the level of a time series. The more sophisticated approach uses the year-over-year change in the variable to eliminate the seasonality in the transformed data series. This eliminates the seasonal component in the mean, and the differenced time series is modeled as a stationary ARMA.

Finally, random walks (also called unit roots) are the most pervasive form of non-stationarity in financial and economic time series. For example, virtually all asset prices behave like a random walk. Directly modeling a time series that contains a unit root is difficult because parameter estimators are biased and are not normally distributed, even in large samples. However, the solution to this problem is simple: model the difference (i.e., $Y_t - Y_{t-1}$).

All non-stationary time series contain trends that may be *deterministic* or *stochastic*. For deterministic trends (e.g., time trends and deterministic seasonalities), knowledge of the period is enough to measure, model, and forecast the trend. On the other hand, *random walks* are the most important example of a stochastic trend. A time series that follows a random walk depends equally on all past shocks.

[1] In some applications, financial data have seasonality-like issues at higher frequencies, for example the day of the week or the hour of the day. While seasonalities are conventionally thought of as applying to annual cycles in data, the concept is applicable to other time series that contain a time-dependent pattern.

The correct approach to modeling time series with trends depends on the source of the non-stationarity. If a time series only contains deterministic trends, then directly capturing the deterministic effects is the best method to model the data. In fact, removing the deterministic trends produces a detrended series that is covariance-stationary.

On the other hand, if a time series contains a stochastic trend, the only transformation that produces a covariance-stationary series is a difference (i.e., $Y_t - Y_{t-j}$ for some value of $j \geq 1$). Differencing a time series also removes time trends and, when the difference is constructed at the seasonal frequency, also removes deterministic seasonalities. Ultimately, many trending economic time series contain both deterministic and stochastic trends, and so differencing is a robust method for removing both effects.

11.1 TIME TRENDS

Polynomial Trends

A time trend deterministically shifts the mean of a series. The most basic example of a non-stationary time series is a process with a linear time trend:

$$Y_t = \delta_0 + \delta_1 t + \epsilon_t, \qquad (11.1)$$

where $\epsilon_t \sim WN(0, \sigma^2)$ and $t = 1, 2, \ldots$.

This process is non-stationary because the mean depends on time:

$$E[Y_t] = \delta_0 + \delta_1 t$$

In this case, the time trend δ_1 measures the average change in Y between subsequent observations.

The linear time trend model generalizes to a polynomial time trend model by including higher powers of time:

$$Y_t = \delta_0 + \delta_1 t + \delta_2 t^2 + \cdots + \delta_m t^m + \epsilon_t \qquad (11.2)$$

In practice, most time trend models are limited to first (i.e., linear) or second-degree (i.e., quadratic) polynomials.

The parameters of a polynomial time trend model can be consistently estimated using OLS. The resulting estimators are asymptotically normally distributed, and standard errors and t-statistics can be used to test hypotheses if *the residuals are white noise*. If the residuals are not compatible with the white noise assumption, then the t-statistics and model R^2 are misleading.

A linear trend model implies that the series grows by a constant quantity each period. In economic and financial time series, this is problematic for two reasons.

1. If the trend is positive, then the growth *rate* of the series (i.e., δ_1/t) will fall over time.

2. If δ is less than zero, then Y_t will eventually become negative. However, this is not plausible for many financial time series (e.g., prices or quantities).

In contrast, constant growth *rates* are more plausible than constant growth in most financial and macroeconomic variables. These growth rates can be examined using a log-linear models that relate the natural log of Y_t to a linear trend. For example:

$$\ln Y_t = \delta_0 + \delta_1 t + \epsilon_t \quad (11.3)$$

This model implies that:

$$E[\ln Y_{t+1} - \ln Y_t] = \delta_1$$

or equivalently:

$$E[\ln(Y_{t+1}/Y_t)] = \delta_1$$

so that the growth rate of Y is constant.

Log-linear models can be extended to higher powers of time. For example, the log-quadratic time trend model:

$$\ln Y_t = \delta_0 + \delta_1 t + \delta_2 t^2 + \epsilon_t \quad (11.4)$$

implies that the growth rate of Y_t is $\delta_1 + 2\delta_2 t$, and so depends on time. In practice, log-linear models are sufficient to capture the trend in series that grow over time.

As an example, consider a data set that measures the real GDP of the United States on a quarterly basis from 1947 until the end of 2018. The specification of the trend models is:

$$RGDP_t = \delta_0 + \delta_1 t + \delta_2 t^2 + \epsilon_t,$$

where $\delta_2 = 0$ in the linear model.

The parameter estimates and t-statistics are reported in the top two rows of Table 11.1. The linear model indicates that real GDP grows by a constant USD 59 billion each quarter, whereas the quadratic model implies that the growth is accelerating.

The top-left panel of Figure 11.1 plots the level of real GDP and the fitted values from the two models. The top-right panel plots the residuals:

$$\hat{\epsilon}_t = RGDP_t - \hat{\delta}_0 - \hat{\delta}_1 t - \hat{\delta}_2 t^2,$$

where $\hat{\delta}_2 = 0$ in the linear model.

Both residual series are highly persistent and not white noise, indicating that these models are inadequate to describe the changing level of real GDP.

The bottom two rows of Table 11.1 contain the estimated parameters and t-statistics from log-linear and log-quadratic models. The log-linear model indicates that real GDP grows by 0.79% per quarter (i.e., about 3% per year). Meanwhile, the log-quadratic indicates that the initial growth rate is higher—1.1% per quarter at the beginning of the sample—and that it is slowly declining. The sample has 288 quarters of data, and so the quarterly growth rate is 0.829%(= 1.1% − 0.00094% × 288) at the end of the sample.

The bottom-left panel of Figure 11.1 plots the natural log of real GDP along with the fitted values from the two models. The right panel shows the residuals from these two regressions. The errors from the log-specification are still clearly not white noise, and so these pure time trend models are also inadequate to explain the growth of real GDP.

The final column of Table 11.1 reports the R^2 of the models. These values are extremely high and the quadratic models both have R^2 above 99%. High R^2 in trending series are inevitable, and the R^2 will approach 100% in all of four specifications as the sample size grows. In practice, R^2 is not considered to be a useful measure of model fit in trending time series. Instead, residual diagnostics or other formal tests are used to assess model adequacy.

Table 11.1 **Estimation Results for Linear, Quadratic, Log-Linear, and Log-Quadratic Models Fitted to US Real GDP. The t-Statistics Are Reported in Parentheses. The Final Column Reports the R^2 of the Model**

Dependent	δ_0	δ_1	δ_2	R^2
RGDP	−43.29	59.38		0.964
	(−0.384)	(87.81)		
RGDP	1966	17.80	0.144	0.996
	(32.69)	(18.53)	(44.68)	
ln(RGDP)	7.712	0.00789		0.990
	(972.5)	(165.9)		
ln(RGDP)	7.581	0.011	−9.4e-06	0.997
	(1278)	(112.0)	(−29.60)	

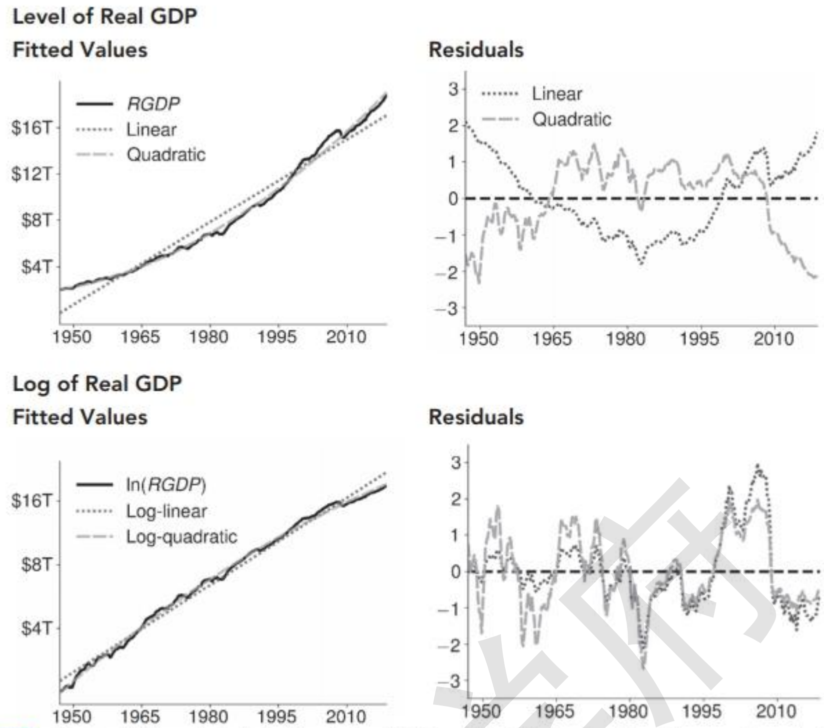


Figure 11.1 **The top-left panel plots the level of real GDP and the fitted values from linear and quadratic trend models. The top-right panel plots the residuals from the linear and quadratic trend models. The bottom-left panel plots the log of real GDP and the fitted values from log-linear and log-quadratic time trend models. The bottom-right plots the residuals from the log-linear and log-quadratic models.**

11.2 SEASONALITY

A seasonal time series exhibits deterministic shifts throughout the year. For example, new housing has a strong seasonality that peaks in the summer months and slows in winter. Meanwhile, natural gas consumption has the opposite seasonal pattern: rising in winter and subsiding in summer. Both seasonalities are driven by predictable events (e.g., the weather).

Seasonalities also appear in financial and economic time series. Many of these seasonalities result from the features of markets or regulatory structures. For example, the January effect posits that equity market returns are larger in January than they are in the other months of the year.[2] Meanwhile, the trading volume of commodity futures varies predictably across the year, reflecting periods when information about crop growing conditions is revealed. Energy prices also have strong seasonalities, although the pattern and strength of the seasonality depend on the geographic location of the market.

Calendar effects generalize seasonalities to cycles that do not necessarily span an entire year. For example, an anomaly has

[2] Tax optimization, where investors sell poorly performing investments in late December to realize losses and then reinvest in January, is one plausible explanation for this phenomenon.

been identified where Monday returns are statistically significantly lower than Friday returns in US equity data.[3] Calendar effects also appear in intra-daily data, where trading volume and price volatility follow a "U" shaped pattern of spikes during the market opening/closing and a mid-day lull.

Seasonal Dummy Variables

Deterministic seasonalities produce differences in the mean of a time series that are simple to model using dummy variables.

Suppose that Y_t is a seasonal time series that has a different mean in each period. The seasonality repeats each s periods (e.g., every four in a quarterly series or 12 in a monthly series). The seasonalities are directly modeled as:

$$Y_t = \delta + \gamma_1 I_{1t} + \gamma_2 I_{2t} + \cdots + \gamma_{s-1} I_{s-1t} + \epsilon_t, \quad (11.5)$$

where

$$I_{jt} = 1 \text{ when } t \,(mod\, s) = j, \text{ and}$$

$$I_{jt} = 0 \text{ when } t \,(mod\, s) \neq j$$

The function $a\,(mod\, b)$ yields the remainder of $\frac{a}{b}$. For example, $10\,(mod\, 6) = 4$.

Note that dummy variable I_{st} is omitted to avoid what is known as the *dummy variable trap*, which occurs when dummy variables are multicollinear.[4]

The mean in period 1 is:

$$E[Y_1] = \delta + \gamma_1$$

The mean in period 2 is:

$$E[Y_2] = \delta + \gamma_2$$

The mean in period s is:

$$E[Y_s] = \delta$$

because all of the included dummy variables are zero and the dummy I_{st} is omitted.

The parameters of this model are estimated using OLS by regressing Y_t on a constant and $s-1$ dummy variables. δ is interpreted as the average of Y_t in period s, whereas γ_j measures the difference between the period j mean and the period s mean.

[3] A recent explanation of this difference posits that key macroeconomic announcements, for example, the number of jobs created in the previous month, are more likely to occur on Friday than on other days, and that the additional return is compensation for risk linked to holding stocks on days when such announcements are made.

[4] For example, consider a model with separate dummy variables for when a light switch turned on (I_{on}) and when it is turned off (I_{off}). These two variables are multicollinear, because if $I_{on} = 1$ then $I_{off} = 0$ (and vice versa). This problem is avoided by simply omitting one of these variables.

Note that an alternative and equally valid approach is to include all s dummy variables but to exclude the intercept term. The model would then be:

$$Y_t = \gamma_1 I_1 t + \gamma_2 I_{2t} + \cdots + \gamma_{s-1} I_{s-1t} + \gamma_s I_{st} + \epsilon_t$$

The parameter interpretations are then slightly different (and simpler) compared with the alternative specification described above. Here, the estimate on dummy j is the average value for Y in that season (e.g., $E[Y_1] = \gamma_1$, $E[Y_2] = \gamma_2, \cdots, E[Y_s] = \gamma_s$, etc). As an example, consider the monthly growth of gasoline consumption in the United States from 1992 until 2018. This time series has a strong seasonality, with consumption being much higher in summer months than in winter months. Figure 11.2 plots the percentage growth rate, which is defined as:

$$100 \times (\ln GC_t - \ln GC_{t-1})$$

The seasonality in this time series is obvious: March is consistently the highest growth month, whereas September is one of four months with large declines in gasoline consumption.

A seasonal model is estimated on these growth rates:

$$\Delta GC_t = \delta + \gamma_1 I_{1t} + \gamma_2 I_{2t} + \cdots + \gamma_{11} I_{11t} + \epsilon_t,$$

where I_{1t} is the January dummy, I_{2t} is a February dummy and so forth.

The fitted values from this model are plotted in the bottom-left panel of Figure 11.2 and the residuals are plotted in the bottom-right panel. While the residuals appear to be approximately white noise, the Ljung-Box Q-test statistic using ten lags is 33. Because this test statistic is distributed χ^2_{10}, the p-value of the test statistic is less than 0.1%, and so the null hypothesis that the residuals are not serial correlated is rejected. Therefore, this model is not adequate to fully describe the dynamics in the data.

11.3 TIME TRENDS, SEASONALITIES, AND CYCLES

Deterministic terms (e.g., time trends and seasonalities) are often insufficient to describe economic time series. As a result, the residuals from models that only contain deterministic terms tend not to be white noise.

When the residuals are not white noise but appear to be covariance-stationary, then the deterministic terms can be combined with AR or MA terms to capture the three components of a time series: the trend, the seasonality, and the cyclicality.

A process is trend-stationary if removing a deterministic time trend produces a series that is covariance-stationary. For example, if the residuals from the model:

$$Y_t = \delta_0 + \delta_1 t + \epsilon_t$$

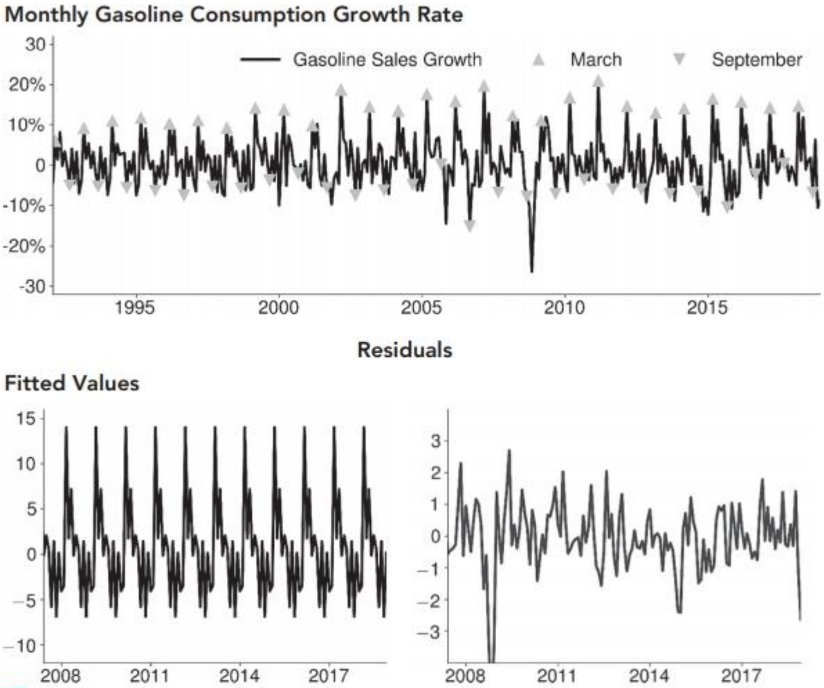


Figure 11.2 The top panel plots the month-over-month growth rate of gasoline consumption in the United States. The upticks indicate the growth rate in March of each year and the downticks mark the growth rate in September. The bottom panel plots the fitted values (left) and residuals (right) from a model with monthly seasonal dummy variables.

are not white noise but are stationary (e.g., follow an AR(1) process), then the model can be augmented to capture the dynamics. This means that adding an AR term to the model should produce residuals that appear to be white noise:

$$Y_t = \delta_0 + \delta_1 t + \phi Y_{t-1} + \epsilon_t$$

If there is a seasonal component to the data, then seasonal dummies can be added as well to produce

$$Y_t = \delta_0 + \delta_1 t + \sum_{i=1}^{s-1} \gamma_i I_{it} + \phi Y_{t-1} + \epsilon_t,$$

where δ_1 captures the long-term trend of the process, γ_i measure seasonal shifts in the mean from the trend growth (i.e., $\delta_1 t$), and ϕY_{t-1} is an AR term that captures the cyclical component.[5]

[5] This example only includes a single AR component but the cyclical component can follow any ARMA(p, q) process.

While adding a cyclical component is a simple method to account for serial correlation in trend-stationary data, many trending economic time series are *not* trend-stationary. If a series contains a unit root (i.e., a random walk component), then detrending cannot eliminate the non-stationarity. Unit roots and trend-stationary processes have different behaviors and properties. Tests of trend stationarity against a random walk are presented in Section 11.4.

The top panel of Figure 11.3 shows the log of gasoline consumption in the United States (i.e., ln GC_t). This series has not been differenced and so the time trend—at least until the financial crisis of 2008—is evident. The predictable seasonal deviations from the time trend are also clearly discernable. Both time trends and seasonalities are deterministic, so that the regressors in the model:

$$Y_t = \delta_0 + \delta_1 t + \delta_2 t^2 + \sum_{i=1}^{11} \gamma_i I_{it} + \epsilon_t$$

only depend on time.

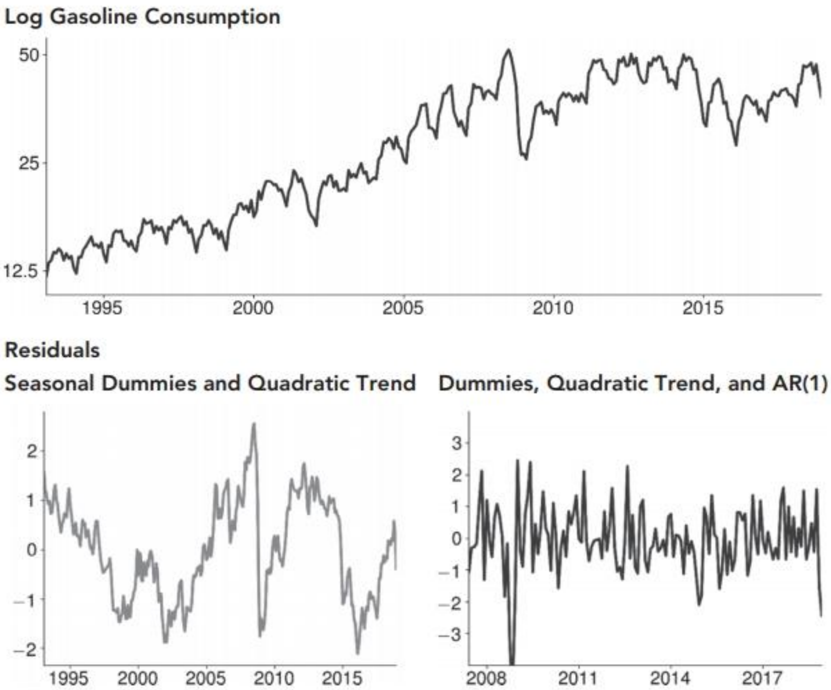


Figure 11.3 **The top panel plots the natural log of gasoline consumption in the United States (ln GC_t) using the data from Figure 11.2. The bottom-left panel plots the residuals from a model that includes seasonal dummy variables and a quadratic time trend. The bottom-right panel plots the residuals from a model with seasonal dummies, a quadratic trend, and a Seasonal AR(1).**

Estimated parameters, *t*-stats, model R^2, and a Ljung-Box Q-statistic using ten lags are reported in Table 11.2. The model in the first row excludes the quadratic term that, when included (second row), has a negative coefficient that appears to be statistically significant.[6]

While both models produce large R^2 values, the trend makes this measure less useful than in a covariance-stationary time series. The bottom-left panel of Figure 11.3 plots the residuals from the model that includes both seasonal dummies and a quadratic time trend. The residuals are highly persistent and clearly not white noise, indicating that the model is badly misspecified.

[6] The usual *t*-stat only has an asymptotic standard normal distribution when the model is correctly specified in the sense that the residuals are white noise. The residuals in these models are serially correlated and so not white noise.

The third row of Table 11.2 reports parameter estimates from a model that extends the dummy-quadratic trend model with a cyclical component [i.e., an AR(1)] and the bottom-right panel of Figure 11.3 plots the residuals from this model. The AR(1) coefficient indicates the series is very persistent, and the coefficient is very close to 1.

However, none of these models appear capable of capturing all of the dynamics in the data, and the Ljung-Box Q-statistic indicates that the null is rejected in all specifications. The estimated parameters in the autoregression are close to one, which suggests that the detrended series is *not* covariance-stationary. This example highlights the challenges of incorporating cyclical components (e.g., autoregressions) with trends and seasonal dummies.

Table 11.2 Parameter Estimates and t-Statistics from Models That Include Combinations of Seasonal Dummies, Linear and Quadratic Trends, and an AR(1) Fitted to Gasoline Sales. The First Two Lines Report Parameters from Models That Include Seasonal Dummies and a Linear or Quadratic Time Trend Component. The Third Row Reports Estimates from a Model that Adds an AR(1) Component to the Results in the Second Row. The Final Two Columns Report the Model R^2 and the Ljung-Box Q-Statistic Using Ten Lags (p-Value in Parentheses)

Seasonal Dummies	δ_1	δ_2	φ_1	R^2	Q_{10}
✓	0.00428 (40.662)			0.851	2193.3 (0.000)
✓	0.00901 (25.120)	−1.4e-05 (−13.549)		0.908	1752.6 (0.000)
✓	0.000437 (2.567)	−8.25e-07 (−2.343)	0.959 (61.307)	0.993	36.4 (0.000)

11.4 RANDOM WALKS AND UNIT ROOTS

Random walks and their generalization, unit root processes, are the third and most important source of non-stationarity in economic time series. A simple random walk process evolves according to:

$$Y_t = Y_{t-1} + \epsilon_t \qquad (11.6)$$

Substituting $Y_{t-1} = Y_{t-2} + \epsilon_{t-1}$ into the random walk:

$$Y_t = (Y_{t-2} + \epsilon_{t-1}) + \epsilon_t$$

shows that Y_t depends on the previous two shocks and the value of the series two periods in the past (i.e., Y_{t-2}).

Repeating the substitution until observation zero is reached gives:

$$Y_t = Y_0 + \sum_{i=1}^{t} \epsilon_i \qquad (11.7)$$

It can be seen that Y_t depends equally on every shock between periods 1 and t, as well as on an initial value Y_0.

Recall that in a stationary AR(1):

$$\begin{aligned} Y_t &= \phi Y_{t-1} + \epsilon_t \\ &= \epsilon_t + \phi\epsilon_{t-1} + \phi^2\epsilon_{t-2} + \cdots + \phi^{t-1}\epsilon_0 + \phi^t Y_0 \end{aligned}$$

Stationarity requires that $|\phi| < 1$ so that for t large enough, both the initial value and shocks in the distant past make a negligible contribution to Y_t. Random walks place equal weight on all shocks and the initial value, and so a shock in period t permanently affects all future values. Unlike covariance-stationary time series, which always mean revert, random walks become more dispersed over time. Specifically, the variance of a random walk is

$$V[Y_t] = t\sigma^2$$

Unit Roots

Unit roots generalize random walks by adding short-run stationary dynamics to the long-run random walk. A unit root process is usually described using a lag polynomial:

$$\begin{aligned} \psi(L)Y_t &= \theta(L)\epsilon_t \\ (1 - L)\phi(L)Y_t &= \theta(L)\epsilon_t, \end{aligned} \qquad (11.8)$$

where $\epsilon_t \sim WN(0, \sigma^2)$ is a white noise process and $\theta(L)\epsilon_t$ is an MA.

For example, the AR(2)

$$Y_t = 1.8Y_{t-1} - 0.8Y_{t-2} + \epsilon_t$$

can be written with a lag polynomial as:

$$\begin{aligned} (1 - 1.8L + 0.8L^2)Y_t &= \epsilon_t \\ (1 - L)(1 - 0.8L)Y_t &= \epsilon_t \end{aligned}$$

The unit root is evident in the factored polynomial and in fact the two roots of this polynomial would be given by solving the characteristic equation: $(1 - z)(1 - 0.8z) = 0$, which gives $z = 1$ and $z = \frac{1}{0.8} = 1.25$. It should be evident that this process is called a unit root precisely because one of the roots in the characteristic equation is equal to unity (i.e., 1). The remaining root is greater than one in absolute value and is therefore a stationary root. When a process contains one stationary root and one non-stationary root, the latter dominates the properties of the series.

The Problems with Unit Roots

Unit roots and covariance-stationary processes have important differences. There are three key challenges when modeling a time series that has a unit root.

1. Parameter estimators in ARMA models fitted to time series containing a unit root are not normally distributed but instead have what is called a Dickey-Fuller (DF) distribution. The DF distribution is asymmetric, sample-size dependent, and its critical values depend on whether the model includes deterministic time trends. In general, the DF distribution also has fatter tails than a normal distribution and so the critical values are larger in absolute value for the DF. These features make it difficult to perform inference and model selection when fitting models on time series that contain unit roots.
2. Ruling out spurious relationships is difficult. A pair of time series have a spurious relationship when there are no fundamental links between them, but a regression of one on the other produces a coefficient estimate that is large and *seemingly* statistically different from zero when using conventional statistical distributions to obtain the critical values. When a time series contains a unit root, it is common to find spurious relationships with other time series that have a time trend or a unit root. Predictive modeling with unit roots is risky because models with highly significant coefficients may be completely misleading and produce large out-of-sample forecasting errors.
3. A unit root process does not mean revert. A stationary AR is mean-reverting, and so the long-run mean can be estimated. Mean-reversion affects the forecasts of stationary AR models so that $E_s[Y_{t+h}] \approx E[Y_s]$ when h is large. In contrast, a random walk is not stationary and $E_s[Y_{t+h}] = Y_t$ for any horizon h. Figure 11.4 plots two simulated AR(2) models $Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \epsilon_t$ using the same initial value and shocks. The models only differ in the coefficients. The stationary AR is parameterized using $\phi_1 = 1.8$ and $\phi_2 = -0.9$. The unit root has coefficients $\phi_1 = 1.8$ and $\phi_2 = -0.8$. The processes track closely for the first 20 observations, but then diverge. The stationary AR process is mean reverting to zero and regularly crosses this value. The random walk is not mean-reverting and so does not have the tendency to "find a level."

The solution to all three problems is to difference a time series that contains a unit root. If Y_t has a unit root, then the difference:

$$\Delta Y_t = Y_t - Y_{t-1}$$

does not.

For example, in a random walk with a drift:

$$Y_t = \delta + Y_{t-1} + \epsilon_t,$$

where $\epsilon_t \sim WN(0, \sigma^2)$, the difference:

$$\Delta Y_t = \delta + Y_{t-1} - Y_{t-1} + \epsilon_t = \delta + \epsilon_t$$

is a constant plus a white noise shock. In the general case of a unit root process:

$$\begin{aligned}(1 - L)\phi(L)Y_t &= \epsilon_t \\ \phi(L)((1 - L)Y_t) &= \epsilon_t \\ \phi(L)\Delta Y_t &= \epsilon_t \end{aligned} \quad (11.9)$$

Because $\phi(L)$ is a lag polynomial of a stationary process, then the random variable defined by the difference (i.e., ΔY_t) must be stationary.

Testing for Unit Roots

Unit root tests examine whether a process contains a unit root. They can also distinguish trend-stationary models from unit root processes with drift.

It is important (but tricky in practice) to correctly distinguish time series that are persistent but stationary from time series that

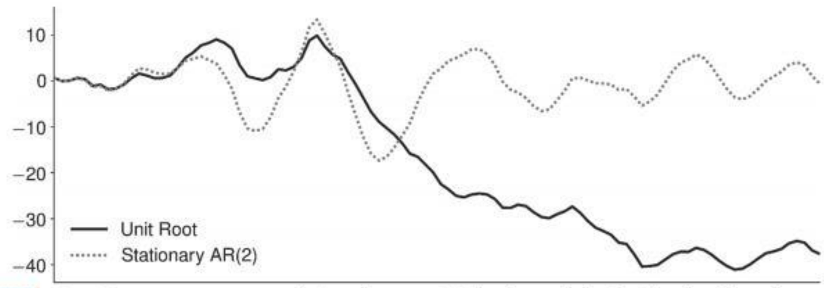


Figure 11.4 Two AR processes generated using the same initial value and shocks. The data from the stationary AR model are generated according to $Y_t = 1.8Y_{t-1} - 0.9Y_{t-2} + \epsilon_t$. The data from the unit root process are generated by $Y_t = 1.8Y_{t-1} - 0.8Y_{t-2} + \epsilon_t$.

contain a unit root. While a unit root can only be removed by differencing, taking the difference of a series that is already stationary is known as over-differencing. An over-differenced time series requires a model that is more complex than would be required to explain the dynamics in the unmodified series.

For example, if:

$$Y_t = \delta + \epsilon_t$$

then

$$\Delta Y_t = \delta + \epsilon_t - \delta - \epsilon_{t-1} = \epsilon_t - \epsilon_{t-1}$$

The original series is a trivial model—an AR(0). The over-differenced series is a MA(1) that is not invertible. The additional complexity in models of over-differenced series increases parameter estimation error and so reduces the accuracy of forecasts.

The Augmented Dickey-Fuller (ADF) specification is the most widely used unit root test. An ADF test is implemented using an OLS regression where the difference of a series is regressed on its lagged *level*, relevant deterministic terms, and lagged differences. The general form of an ADF regression is

$$\Delta Y_t = \underbrace{\gamma Y_{t-1}}_{\text{Lagged Level}} + \underbrace{\delta_0 + \delta_1 t}_{\text{Deterministic}} + \underbrace{\lambda_1 \Delta Y_{t-1} + \ldots + \lambda_p \Delta Y_{t-p}}_{\text{Lagged Differences}} \quad (11.10)$$

The ADF test statistic is the *t*-statistic of $\hat{\gamma}$. To understand the ADF test, consider a implementing a test with a model that only includes the lagged level:

$$\Delta Y_t = \gamma Y_{t-1} + \epsilon_t$$

This specification would be known as a DF test, so the "augmented" term in the ADF test refers to the addition of lagged differenced terms in the test equation. If Y_t is a random walk, then:

$$\begin{aligned} Y_t &= Y_{t-1} + \epsilon_t \\ Y_t - Y_{t-1} &= Y_{t-1} - Y_{t-1} + \epsilon_t \\ \Delta Y_t &= 0 \times Y_{t-1} + \epsilon_t \end{aligned}$$

so that the value of γ is 0 when the process is a random walk. Under the null H_0: $\gamma = 0$, Y_t is a random walk and so is non-stationary. The alternative is H_1: $\gamma < 0$, which corresponds to the case that Y_t is covariance-stationary. Note that the alternative is one-sided, and the null *is not* rejected if $\gamma > 0$. Positive values of γ correspond to an AR coefficient that is larger than 1, and so the process is explosive and not covariance-stationary.

Implementing an ADF test on a time series requires making two choices: which deterministic terms to include and the number of lags of the differenced data to use. The number of lags to include is simple to determine—it should be large enough to absorb any short-run dynamics in the difference ΔY_t.

The lagged differences in the ADF test are included to ensure that $\hat{\epsilon}_t$ is a white noise process. The recommended method to select the number of lagged differences is to choose the lag length to minimize AIC. The maximum lag length should be set to a reasonable value that depends on the length of the time series and the frequency of sampling (e.g., 24 for monthly data or eight for quarterly data).

Recall that the AIC tends to select a larger model than criteria such as the BIC. This approach to selecting the lag length is preferred because it is essential that the residuals are approximately white noise otherwise the unit root test statistic will not follow the required DF or ADF distribution, and so selecting too many lags is better than selecting too few. Ultimately, any reasonable lag length selection procedure—IC-based, graphical, or general-to-specific selection—should produce valid test statistics and the same conclusion.

The included deterministic terms have a more significant impact on the ADF test statistic. The DF distribution depends on the choice of deterministic terms. Including more terms skews the distribution to the left, and so the critical value becomes more negative as additional deterministic terms are included. For example, the (one-sided) 5% critical values in a time series with 250 observations are −1.94 when no deterministic terms are included, −2.87 when a constant is included, and −3.43 when a constant and trend are included. All things equal, adding additional deterministic terms makes rejecting the null more difficult when a time series *does not* contain a unit root. This reduction in the power of an ADF test suggests a conservative approach when deciding which deterministic trends to include in the test.

On the other hand, if the time series is trend-stationary, then the ADF test must include a constant. If the ADF regression is estimated without the constant, then the null is asymptotically *never rejected*, and the power of the test is zero.

Avoiding this outcome requires including any relevant deterministic terms. The recommended method to determine the relevant deterministic terms is to use *t*-statistics to test their statistical significance using a size of 10%. Any deterministic regressor that is statistically significant at the 10% level should be included. If the trend is insignificant at the 10% level, then it can be dropped, and the ADF test can be rerun including only a constant. If the constant is also insignificant, then it too can be dropped, and the test rerun with no deterministic components. However, most applications to financial and macroeconomic time series require the constant to be included.

When the null of a unit root cannot be rejected, the series should be differenced. The best practice is to repeat the ADF test on the differenced data to ensure that it is stationary. If the difference is also non-stationary (i.e., the null cannot be rejected on the difference), then the series should be double differenced. If the double-differenced data are not stationary, then this is an indication that some other transformation may be

required before testing stationarity. For example, if the series is always positive, it is possible that the natural log should be used instead of the unadjusted data.

In practice, research has identified that the majority of economic and financial time series contain one unit root—that is, they can be made stationary by applying the differencing operator once. We would then say that the original unit root series are "integrated of order one" (I(1)) and the differenced series are I(0). There is only one class of series for which there is some evidence that it may contain two unit roots and thus might be I(2): consumer price series such as the CPI or RPI. If this is the case—and the evidence is mixed—the series would require differencing twice to make a stationary series.

Table 11.3 contains the results of six ADF tests on the natural log of real GDP. The top panel contains tests of log RGDP, and the bottom panel contains tests on the growth rate (i.e., the difference of the log).

Each panel contains ADF tests for all three configurations of deterministic terms. In the top panel, the trend is not statistically significant at the 10% level because its absolute *t*-statistic is less than 1.645. The lack of statistical significance indicates that the trend should be dropped and the test run including only a constant, which is shown in the middle row. The constant is highly significant, and so this model is the preferred specification.

The *t*-statistic on the coefficient on the lagged level γ is -2.330. Using the critical values listed in the right columns, the *t*-statistic is above (i.e., less negative than) the 5% critical value, and the null is not rejected. The top row excludes deterministic terms and produces a positive estimate of γ. This is a common pattern—a positive estimate of γ usually indicates an inappropriate choice of deterministic terms. The number of lags is chosen by minimizing the AIC across all models that include up to eight lags.

The bottom panel contains the same three specifications of deterministic terms on the growth rate of real GDP (i.e., $\Delta \ln RGDP_t$). In this case, the time trend is statistically significant, and so this model should be used to test for a unit root. The *t*-statistic on γ is -8.165, which is below (i.e., more negative than) the critical value of -3.426 for a 5% test, and so the

Table 11.3 **ADF Test Results for the Natural Log of Real GDP (top panel) and the Difference of the Log of Real GDP (i.e., GDP Growth, bottom panel). The Column γ Reports the Parameter Estimate from the ADF Test Regression and the ADF Statistic in Parentheses. The Columns δ_0 and δ_1 Report Estimates of the Constant and Linear Trend Along with *t*-Statistics. The Lags Column Reports the Number of Lags in the ADF Regression as Selected by the AIC. The Final Columns Report the 5% and 1% Critical Values and Test Statistics Less (i.e., More Negative) Than the Critical Value that Leads to Rejection of the Null Hypothesis of a Unit Root**

Log of Real GDP						
Deterministic	γ	δ_0	δ_1	**Lags**	**5% CV**	**1% CV**
None	5.14e-04 (5.994)			3	−1.942	−2.574
Constant	−1.92e-03 (−2.330)	0.023 (3.053)		5	−2.872	−3.454
Trend	−8.27e-03 (−1.018)	0.072 (1.147)	5.09e-05 (0.786)	5	−3.426	−3.991
Growth of Real GDP						
Deterministic	γ	δ_0	δ_1	**Lags**	**5% CV**	**1% CV**
None	−0.133 (−2.184)			8	−1.942	−2.574
Constant	−0.626 (−8.414)	4.87e-03 (6.287)		2	−2.872	−3.454
Trend	−0.783 (−8.165)	8.13e-03 (5.641)	−1.47e-05 (−2.237)	4	−3.426	−3.991

Table 11.4 ADF Test Results for the Default Premium, Defined as the Difference between the Interest Rates on Portfolios of Aaa- and Baa-Rated Bonds. The Column γ Reports the Parameter Estimate from the ADF and the ADF Statistic in Parentheses. The Columns δ_0 and δ_1 Report Estimates of the Constant and Linear Trend along with t-Statistics. The Lags Column Reports the Number of Lags in the ADF Regression as Selected by the AIC. The Final Columns Report the 5% and 1% Critical Values

Default Premium						
Deterministic	γ	δ_0	δ_1	**Lags**	**5% CV**	**1% CV**
None	−5.34e-03 (−1.220)			16	−1.942	−2.571
Constant	−0.042 (−3.251)	0.045 (3.037)		10	−2.868	−3.444
Trend	−0.048 (−3.473)	0.063 (3.009)	−5.00e-05 (−1.217)	10	−3.420	−3.978

test indicates that the null of a unit root in the differenced series should be rejected and so we conclude that the real GDP growth rate does not contain a unit root. The ADF test that only includes a constant produces a similar conclusion. Excluding both deterministic terms produces an ADF test statistic that can be rejected at 5% but not 1%. This value is not reliable because the deterministic terms are statistically significant and so need to be included for a valid test.

Table 11.4 shows the results of testing the default premium—defined as the difference between the yield on a portfolio of Aaa-rated corporate bonds and Baa-rated corporate bonds. The tests are implemented using 40 years of monthly data between 1979 and 2018.

For this data, the time trend is not statistically significant and so can safely be dropped. The null that the constant is zero, however, is explicitly rejected at the 10% level, and so the ADF test should be evaluated using a specification that includes a constant. The t-statistic on γ is −3.23, which is less than the 5% (but not the 1%) critical value. In the previous chapter, we saw that the default premium is highly persistent. These ADF tests indicate that it is also covariance-stationary. The number of lags is chosen to minimize the AIC across all models that include between zero and 24 lags.

Seasonal Differencing

Seasonal differencing is an alternative approach to modeling seasonal time series with unit roots. The seasonal difference is constructed by subtracting the value in the *same period in the previous year*. This transformation eliminates deterministic seasonalities, time trends, *and* unit roots.

For example, suppose Y_t is a quarterly time series with both deterministic seasonalities and a non-zero growth rate, so that:

$$Y_t = \delta_0 + \gamma_1 I_{1t} + \gamma_2 I_{2t} + \gamma_3 I_{3t} + \delta_1 t + \epsilon_t,$$

where $\epsilon_t \sim WN(0, \sigma^2)$.

The seasonal difference is denoted $\Delta_4 Y_t$ and is defined $Y_t - Y_{t-4}$. Substituting in the difference and simplifying, this difference evolves according to:

$$Y_t - Y_{t-4} = \delta_1(t - (t-4)) + \epsilon_t - \epsilon_{t-4}$$
$$\Delta_4 Y_t = 4\delta_1 + \epsilon_t - \epsilon_{t-4}$$

The seasonal dummies are eliminated because:

$$\gamma_j I_{jt} - \gamma_j I_{jt-4} = \gamma_j(I_{jt} - I_{jt-4}) = \gamma_j \times 0$$

This is because $I_{jt} = I_{jt-4}$ for any value of t (because the observations are four periods apart).

The differenced series is a MA(1), and so is covariance-stationary.

By removing trends, seasonality, and unit roots, seasonal differencing produces a time series that is covariance-stationary. The seasonally differenced time series can then be modeled using standard ARMA models. Seasonally differenced series are directly interpretable as the year-over-year change in Y_t or, when logged, as the year-over-year growth rate in Y_t.

Figure 11.5 contains a plot of the seasonal difference in the log gasoline consumption series. This series is constructed by subtracting log consumption in the same month of the previous year, $100 \times (\ln GC_t - \ln GC_{t-12})$.

Year-over-year Growth in Gasoline Consumption

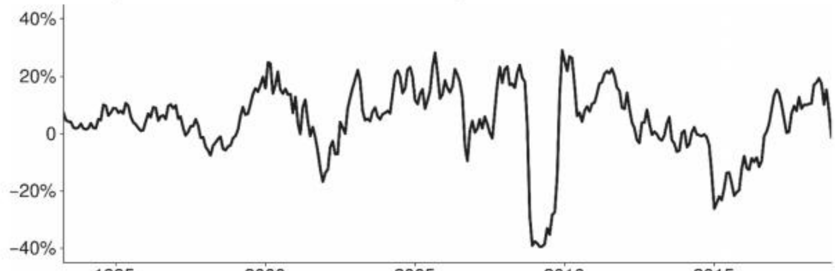


Figure 11.5 **The seasonal difference, or year-over-year growth, of log gasoline consumption in the United States, $100 \times (\ln GC_t - \ln GC_{t-12})$.**

The difference eliminates the striking seasonal pattern present in the month-over-month growth rate plotted in Figure 11.2, and so this series can be modeled as a covariance-stationary time series without dummies.

11.5 SPURIOUS REGRESSION

Spurious regression is a common pitfall when modeling the relationship between two or more non-stationary time series.

For example, suppose Y_t and X_t are independent unit roots with Gaussian white noise shocks, and the cross-sectional regression:

$$Y_t = \alpha + \beta X_t + \epsilon_t$$

is estimated using 100 observations.

When using simulated data, the t-statistic on β is statistically different from zero in over 75% of the simulations when using a 5% test. The rejection rate should be 5%, and the actual rejection rate is extremely high considering that the series are *independent* and so have no relationship. Furthermore, the rejection probability converges to one in large samples, and so a naïve statistical analysis always indicates a strong relationship between the two series when in fact there is none. This phenomenon is called a spurious regression.

Spurious regression is a frequently occurring problem when both the dependent and one or more independent variables are non-stationary. However, it is not an issue in stationary time series, and so ensuring that both X_t and Y_t are stationary (and differencing if not) prevents this problem.

As another example, consider the weekly log of the Russell 1000 index and the log of the Japanese JPY/USD exchange rate.

The current value of the index is regressed on the lag of the exchange rate:

$$\ln R\,1000_t = \alpha + \beta \ln JPYUSD_{t-1} + \epsilon_t$$

The two series are plotted in the top panel of Figure 11.6. The Russell 1000 has an obvious time trend over the sample. The JPY/USD rate may also have a time trend, although it is less pronounced.

The first line of Table 11.5 contains the estimated parameters, t-statistics and R^2 of the model. This naïve analysis suggests that the exchange rate in the previous week is a strong predictor of the price of the Russell 1000 at the end of the subsequent week. This result, also plotted in Figure 11.6, is spurious because both series have unit roots.

The bottom panel of Figure 11.6 plots the residuals (i.e., $\hat{\epsilon}_t$) from this regression. These residuals are highly persistent. Furthermore, conducting an ADF test results in a failure to reject the null that the residuals contain a unit root.

The correct approach uses the differences of the series (i.e., the returns) to explore whether the JPY/USD rate has predictive power. The well-specified regression is

$$\Delta \ln R\,1000_t = \alpha + \beta \Delta \ln JPYUSD_{t-1} + \epsilon_t$$

The differenced data series are both stationary, a requirement for OLS estimators to be well behaved in large samples.

Results from this regression are reported in the second line of Table 11.5. The R^2 is only 0.3%, indicating that the JPY/USD return has little predictive power for the future return of the equity index. This is what we would have expected and we can apply the "spurious regression" term to the levels model since there was no strong theoretical basis for expecting the two time series to have a relationship.

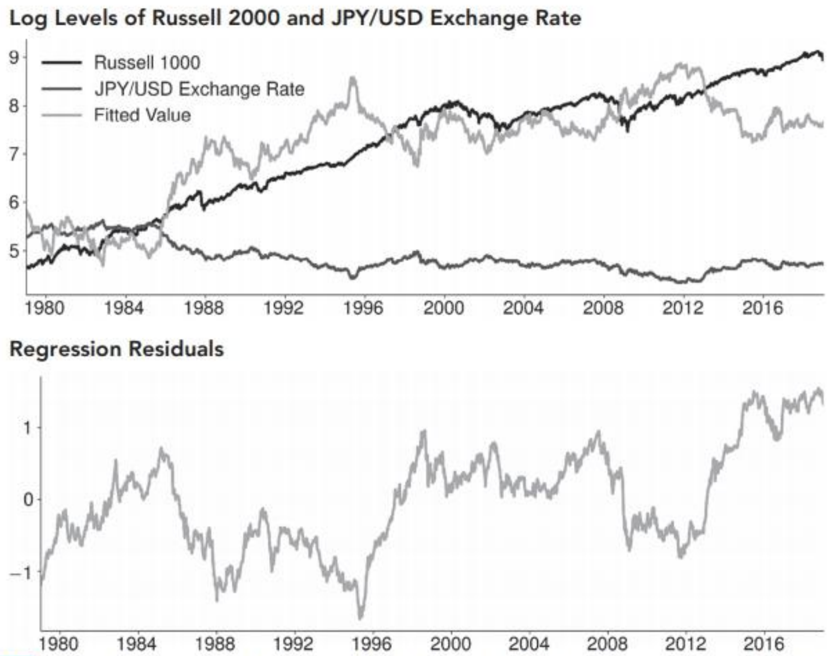


Figure 11.6 The top panel plots the logs of the Russell 1000 index, the JPY/USD exchange rate, and the fitted value from regressing the Russell 1000 on the JPY/USD exchange rate. The bottom panel plots the regression residuals from a spurious regression of the Russell 1000 on the JPY/USD exchange rate.

Table 11.5 Estimation Results from a Spurious Regression (First Row) and a Well-Specified Regression. The First Row Contains Estimation Results from a Model That Regresses the End-of-Week Log of the Russell 1000 Index on the Log JPY/USD Exchange Rate at the End of the Previous Week. The Second Row Shows Estimates from a Model That Regresses the Weekly Return of the Russell 1000 on the Previous Week's JPY/USD Exchange Rate Return

	α	β	R^2
Logs	22.855	−3.232	67.6%
	(95.823)	(−65.932)	
Returns	0.000	−0.055	0.3%
	(4.489)	(−2.414)	

11.6 WHEN TO DIFFERENCE?

Many financial and economic time series are highly persistent but stationary, and differencing is only required when a time series contains a unit root. When a time series cannot be easily categorized as stationary or non-stationary, it is good practice to build models for the series in both levels and differences.

As an example, consider the default premium, which is a highly persistent time series that rejects the null of a unit root when tested.The default premium can be modeled using both approaches: An AR(3) is estimated on the levels, and an AR(2) is estimated on the differences.[7]

[7] Modeling the differences assumes that the process contains a unit root and so the AR order is reduced by one.

Table 11.6 **The First Row Reports Parameter Estimates and t-Statistics of an AR(3) Estimated on the Level of the Default Premium. The Second Row Reports Estimates from an AR(2) Estimated on the Difference in the Default Premium. The Final Row Reports the Estimated Parameters in the Implied AR(3) Reconstructed from the AR(2) Parameters Estimated on the Differences**

	δ	φ_1	φ_2	φ_3
Levels	0.044	1.330	−0.524	0.152
	(0.362)	(29.539)	(−7.307)	(3.385)
Differences	−0.000	0.360	−0.182	
	(−0.014)	(8.028)	(−4.057)	
Diff. as Levels	−0.000	1.360	−0.542	0.182
	(−0.014)	(30.305)	(−7.481)	(4.057)

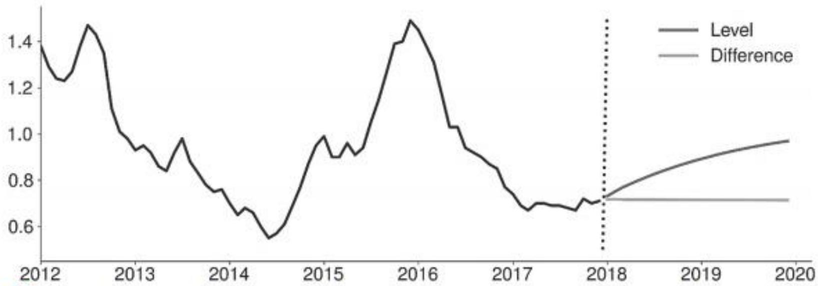


Figure 11.7 **The paths of the forecasts of the default premium starting in January 2018 for models estimated in levels and differences.**

The estimated parameters from these two models are reported in Table 11.6. The model of the differences can be equivalently expressed as a model in levels because:

$$\Delta Y_t = \delta + \phi_1 \Delta Y_{t-1} + \phi_2 \Delta Y_{t-2} + \epsilon_t$$
$$Y_t - Y_{t-1} = \delta + \phi_1(Y_{t-1} - Y_{t-2}) + \phi_2(Y_{t-2} - Y_{t-3}) + \epsilon_t$$
$$Y_t = \delta + (1 + \phi_1)Y_{t-1} + (\phi_2 - \phi_1)Y_{t-2} - \phi_2 Y_{t-3} + \epsilon_t$$

The transformed parameters are reported in the bottom row of Table 11.6. While these parameters appear to be very similar, they imply different forecast paths: The model estimated in levels is mean reverting to 1.06%, whereas the model estimated using differences has short-term dynamics but is not mean reverting (because it contains a unit root).

Figure 11.7 shows the default premium for the four years prior to 2018 and two forecasts for the following 24 months. Whereas the mean reversion in the levels-estimated model is apparent, the differences-estimated model predicts little change in the default premium over the next two years despite having coefficients different from zero. Ultimately, determining which approach works better is an empirical question, and the best practice is to consider models in both levels and differences when series are highly persistent.

11.7 FORECASTING

Constructing forecasts from models with time trends, seasonalities, and cyclical components is no different from constructing forecasts from stationary ARMA models. The forecast is the time T expected value of Y_{T+h}. For example, in a linear time trend model:

$$Y_T = \delta_0 + \delta_1 T + \epsilon_t,$$

so that Y_{T+h} is:

$$Y_{T+h} = \delta_0 + \delta_1(T + h) + \epsilon_{T+h}$$

The time T expectation of Y_{T+h} is:

$$E_T[Y_{T+h}] = \delta_0 + \delta_1(T + h) + E_T[\epsilon_{T+h}]$$
$$= \delta_0 + \delta_1(T + h)$$

because ϵ_{T+h} is a white noise shock and so has a mean of zero.

Forecasting seasonal time series in models with dummies requires tracking the period of the forecast. In the dummy-only model for data with a period of s:

$$Y_t = \delta + \sum_{j=1}^{s} \gamma_j I_{jt} + \epsilon_t$$

the 1-step ahead forecast is:

$$E_T[Y_{T+1}] = \delta + \gamma_j,$$

where $j = (T + 1) \pmod{s}$ is the period of the forecast and the coefficient on the omitted period is $\gamma_0 = 0$.

For example in a quarterly model that excludes the dummy for Q4, if T is period 125, then $E_T[Y_{T+1}] = \delta + \gamma_{(125+1)(\text{mod } 4)} = \delta + \gamma_2$. The h-step ahead forecasts are constructed using an identical process only tracking the period of $T + h$, so that:

$$E_T[Y_{T+h}] = \delta + \gamma_j,$$

where $j = (T + h) \pmod{s}$.

Forecasts from models that include both cyclical components and seasonal dummies are constructed recursively. The procedure is virtually identical to forecasting a stationary ARMA model, and the only difference is that the deterministic term in the h-step forecasts depends on the period of the forecast, $T + h$.

The effect of the recent value diminishes as the horizon grows, and so models containing a cyclical component and with seasonal dummies exhibit mean reversion. However, the mean reversion is *not* to a single unconditional mean as in a covariance-stationary AR(p). Instead, the forecasts from a model with seasonal dummies mean revert to the period-specific means captured by the dummies.

Forecast Confidence Intervals

In some applications, it is useful to construct both the forecasts and confidence intervals to express the uncertainty of the future value. The forecast confidence interval depends on the variance of the forecast error, which is defined as the difference between the time $T + h$ realization and its forecast:

$$Y_{T+h} - E_T[Y_{T+h}]$$

For example, consider a simple linear time trend model:

$$Y_{T+h} = \delta_0 + \delta_1(T + h) + \epsilon_{T+h}$$

In this case:

$$E_T[Y_{T+h}] = \delta_0 + \delta_1(T + h)$$

and the forecast error is ϵ_{T+h}.

When the error is Gaussian white noise $N(0, \sigma^2)$, then the 95% confidence interval for the future value is $E_T[Y_{T+h}] \pm 1.96\sigma$. In practice, σ is not known. However, it can be estimated as the square root of the residual variance from the estimated regression.

Confidence intervals can be constructed for forecasts from any model, whether the process is stationary or not. The confidence interval only depends on the variance of $Y_{T+h} - E_T[Y_{T+h}]$. In models that include an AR or MA component, the forecast error variance depends on both the noise variance as well as the AR and MA parameters.

FORECASTING LEVELS USING LOG MODELS

Using a log transformation complicates forecasting. The log is a nonlinear transformation, and so $E[Y_{T+h}] \neq \exp E[\ln Y_{T+h}]$.

This difference follows from Jensen's inequality because the log is a concave transformation. If the assumption on the noise is strengthened so that it is also Gaussian, $\epsilon_t \stackrel{iid}{\sim} N(0, \sigma^2)$, then the properties of a log-normal can be used to construct forecasts of the level from a model of the log. Recall that if $X \sim N(\mu, \sigma^2)$ then $W = \exp(X) \sim Log\ N(\mu, \sigma^2)$. The mean of W depends on both μ and σ^2, and $E[W] = \exp(\mu + \sigma^2/2)$. For example, in the log-linear trend model:

$$\ln Y_{T+h} = \delta_0 + \delta_1(T + h) + \epsilon_{T+h}$$

the mean is $E_T[\ln Y_{T+h}] = \delta_0 + \delta_1(T + h)$ and σ^2 is the variance of the shock, ϵ_{T+h}, so that $\ln Y_{T+h} \sim N(\delta_0 + \delta_1(T + h), \sigma^2)$. The expected value of Y_{T+h} is then $\exp(\delta_0 + \delta_1(T + h) + \sigma^2/2)$.

11.8 SUMMARY

Non-stationary time series can contain deterministic or stochastic trends. The correct approach to modeling a non-stationary time series depends on the type of trend. If a time series has only deterministic trends (e.g., time trends or deterministic seasonalities), then the trends can be directly modeled in conjunction with the cyclical component. When the series contains a stochastic trend (e.g., a unit root), then the series must be differenced to remove the trend.

Non-stationary time series are challenging to model for three reasons. First, the distribution of parameter estimators depends on whether the time series contains a unit root, and so correctly interpreting test statistics is challenging. Second, it is easy to find spurious relationships that can arise due to deterministic or stochastic trends. Finally, forecasts of non-stationary time series do not mean-revert to a fixed level. Forecasts produced from models built using over- or under-differenced data are likely to perform very poorly, especially at longer horizons.

The Augmented Dickey-Fuller (ADF) test is used to differentiate between a unit root and a trend-stationary process. Running an ADF test is the first step in building a model for a time series that might be non-stationary. If the ADF fails to reject the null that the series contains a unit root, then the series should be differenced. In most cases, the first difference eliminates both the stochastic trend and a linear time trend, if present, and so produces a stationary time series. However, in a seasonal time series, differencing should be implemented at the seasonal frequency. This eliminates deterministic seasonalities, linear time trends, and unit roots.

The following questions are intended to help candidates understand the material. They are *not* actual FRM exam questions.

QUESTIONS

Short Concept Questions

11.1 When modeling $\ln Y_t$ using a time trend model, what is the relationship between $\exp E_T[\ln Y_{T+h}]$ and $E_T[Y_{T+h}]$ for any forecasting period h? Are these ever the same? Assume that the error terms are normally distributed around a mean of zero.

11.2 Suppose that an hourly time series has a calendar effect where the hour of the day matters. How would the dummy variable approach be implemented to capture this calendar effect? How could differencing be used instead to remove the seasonality?

11.3 Why does a unit root with a time trend, $Y_t = \delta_1 + Y_{t-1} + \epsilon_t$ not depend explicitly on t?

11.4 What are the consequences of excluding a deterministic term in an ADF test if it is needed? What are the consequences of including an extra deterministic term that is not required to fit the data?

11.5 Why doesn't the ADF test reject if the t-statistic is large and positive?

Practice Questions

11.6 A linear time trend model is estimated on annual real euro-area GDP, measured in billions of 2010 euros, using data from 1995 until 2018. The estimated model is $RGDP_t = -234178.8 + 121.3 \times t + \hat{\epsilon}_t$. The estimate of the residual standard deviation is $\hat{\sigma} = 262.8$. Construct point forecasts and 95% confidence intervals (assuming Gaussian white noise errors) for the next three years. Note that t is the year, so that in the first observation, $t = 1995$, and in the last, $t = 2018$.

11.7 A log-linear trend model is also estimated on annual euro-area GDP for the same period. The estimated model is $\ln RGDP_t = -18.15 + 0.0136\, t + \hat{\epsilon}_t$, and the estimated standard deviation of ϵ_t is 0.0322. Assuming that the shocks are normally distributed, what are the point forecasts of *GDP* for the next three years? How do these compare with those from in the previous problem?

11.8 The seasonal dummy model $Y_t = \delta + \sum_{j=1}^{3} \gamma_j I_{jt} + \epsilon_t$ is estimated on the quarterly growth rate of housing starts, and the estimated parameters are $\hat{\gamma}_1 = 6.23$, $\hat{\gamma}_2 = 56.77$, $\hat{\gamma}_3 = 10.61$, and $\hat{\delta} = -15.79$ using data until the end of 2018. What are the forecast growth rates for the four quarters of 2019?

11.9 ADF tests are conducted on the log of the ten-year US government bond interest rate using data from 1988 until the end of 2017. The results of the ADF with different configurations of the deterministic terms are reported in the table below. The final three columns report the number of lags included in the test as selected using the AIC and the 5% and 1% critical values that are appropriate for the sample size and included deterministic terms. Do interest rates contain a unit root?

Deterministic	γ	δ_0	δ_1	Lags	5% CV	1% CV
None	−0.003 (−1.666)			7	−1.942	−2.572
Constant	−0.009 (−1.425)	0.010 (1.027)		4	−2.870	−3.449
Trend	−0.085 (−4.378)	0.188 (4.260)	−0.000 (−4.109)	3	−3.423	−3.984

The following questions are intended to help candidates understand the material. They are *not* actual FRM exam questions.

11.10 An AR(2) process is presented:

$$Y_t = kY_{t-1} + mY_{t-2} + \epsilon_t$$

It is determined that the equation has a unit root.

a. What conditions must k and m satisfy for the first difference series to be stable?

b. Express the equation as an AR(1) process.

11.11 The following data are to be evaluated for whether they follow a random walk:

t	Y_t
1	−1.81
2	−2.31
3	−1.65
4	−0.81
5	−0.31
6	−0.37
7	0.47
8	1.10
9	0.53
10	0.47
11	1.00
12	1.68
13	0.90
14	1.99
15	1.31
16	0.76
17	0.43
18	−1.01
19	0.27
20	−2.21
21	−3.21
22	−1.90
23	−0.53
24	−0.60
25	0.34

The specific Dickey Fuller (DF) regression to be evaluated is:

$$y_t - y_{t-1} = my_{t-1} + \epsilon_t$$

The relevant critical values for the DF test are 1%: −2.60, 5%: −1.95.

a. What is the appropriate null hypothesis and alternative hypothesis if a one-sided DF test is employed?

b. Calculate the DF t-statistic, taking $Y_0 = 0$.

Hint: For a regression with no intercept, $y_i = mx_i + \epsilon_i$;

$$m = \frac{\Sigma x_i y_i}{\Sigma x_i^2}$$

ANSWERS

Short Concept Questions

11.1 A time trend model for $\ln Y_t$ can be stated as:

$$\ln Y_t = g(t) + \epsilon_t, \quad \epsilon \sim N(0, \sigma^2),$$

where $g(t)$ is a function of t.

So:

$$E_T[\ln Y_{T+h}] = g(T+h),$$

which gives:

$$\exp E_T[\ln Y_{T+h}] = \exp[g(T+h)]$$

On the other hand:

$$E_T[Y_{T+h}] = E_T[\exp(g(T+h) + \epsilon_{T+h})] = \exp(g(T+h) + E_T[\exp \epsilon_{T+h})],$$

which equals:

$$E_T[Y_{T+h}] = \exp[g(T+h)] + \frac{\sigma^2}{2}$$

And so:

$$E_T[Y_{T+h}] = \exp E_T[\ln Y_{T+h}] + \frac{\sigma^2}{2}$$

These will be equal if the variance is zero (in other words, if the process is completely deterministic.

11.2 Let $s = 24$ represent the hour of the day using a 24-hour clock (e.g. 13 = 1 p.m.). Then:

$$Y_t = g(t) + \gamma_1 I_{1t} + \gamma_2 I_{2t} + \cdots + \gamma_{23} I_{23t} + \epsilon_t,$$

where $I_{jt} = 1$ when $t(mod\ 24) = j$, and $I_{jt} = 0$ when $t(mod\ 24) \neq j$

Differencing this series can be achieved by looking at observations 24 periods (hours) apart from each other (the following presumes that the error terms are iid and normal):

$$Y_{t+24} - Y_t = g(t+24) - g(t) + \epsilon_{t+24} - \epsilon_t$$

Once the deterministic time trend is removed, the remaining part of the series is a covariance-stationary MA(1) process.

11.3 The time trend becomes apparent as the series is propagated backwards:

$$Y_t = \delta_1 + Y_{t-1} + \epsilon_t = 2\delta_1 + Y_{t-2} + \epsilon_t + \epsilon_{t-1} = \cdots$$

$$= t\delta_1 + Y_0 + \sum_{i=1}^{t} \epsilon_i$$

11.4 Excluding required deterministic terms may lead to biased and inconsistent DF test statistics. On the other hand, including superfluous deterministic terms skews the distribution to the left and reduces the power of the test.

11.5 The ADF focuses on the left side of the distribution and rejection occurs if the test statistic is to the left of the demarcation point (e.g., large and negative). A large and positive test statistic would imply an explosive process, not a stationary one as required to reject the null hypothesis.

Solved Problems

11.6 Note that there is no AR or MA component, so the variance remains constant. Therefore, the 95% confidence interval is +/− 1.96*262.8 = +/− 515.1 about the expected value.

As for the expected means:

$$E[RGDP_{2019}] = -234178.8 + 121.3 \times 2019 = 10{,}725.9$$

$$E[RGDP_{2020}] = -234178.8 + 121.3 \times 2020 = 10{,}847.2$$

$$E[RGDP_{2021}] = -234178.8 + 121.3 \times 2021 = 10{,}968.5$$

11.7 In this case:

$$E_T[Y_T] = \exp\left(E_T[\ln Y_T] + \frac{\sigma^2}{2}\right)$$

And the error bounds on the ln are +/− 1.96*0.0322, so the bounds are given in proportional terms rather than fixed values.

$$\text{Bounds_Multiplier} = \exp(\pm 1.96 * 0.0322)$$
$$= \exp(\pm 0.0631) = 0.939, 1.065$$

Calculating $E[\ln RGDP_t]$:

$$E[\ln RGDP_{2019}] = -18.15 + 0.0136 * 2019 = 9.308$$

$$E[\ln RGDP_{2020}] = -18.15 + 0.0136 * 2020 = 9.322$$

$$E[\ln RGDP_{2021}] = -18.15 + 0.0136 * 2020 = 9.336$$

Furthermore,

$$\frac{\sigma^2}{2} = \frac{0.0322^2}{2} = 0.0005$$

(which will only make a small impact in this example)

So:

$$E[RGDP_{2019}] = \exp(9.308 + 0.0005) = 11{,}031.4$$

$$E[RGDP_{2020}] = \exp(9.322 + 0.0005) = 11{,}186.9$$

$$E[RGDP_{2021}] = \exp(9.336 + 0.0005) = 11{,}344.6$$

And the 95% confidence bands are given as:

$$95\%_{CB_{RGDP_{2019}}} = [0.939 * 11031.4, 1.065 * 11031.4]$$

$$= [10{,}358.5, 11{,}748.4]$$

$$95\%_{CB_{RGDP_{2020}}} = [10{,}504.5, 11{,}914.0]$$

$$95\%_{CB_{RGDP_{2021}}} = [10{,}652.6, 12{,}082.0]$$

In comparison with #1, the bands are growing in size and overall the results are a bit bigger.

11.8 Though there is variance from quarter-to-quarter, the expected value of Y_t is the same for any two observations of the same quarter, regardless of the year.

Accordingly:

$$E[Y_{Q1}] = \delta + \sum_{j=1}^{3} \gamma_j I_{jt} = -15.79 + 6.23 * 1 + 56.77 * 0 + 10.61 * 0 = -9.56$$

$$E[Y_{Q2}] = \delta + \sum_{j=1}^{3} \gamma_j I_{jt} = -15.79 + 6.23 * 0 + 56.77 * 1 + 10.61 * 0 = 40.98$$

$$E[Y_{Q3}] = \delta + \sum_{j=1}^{3} \gamma_j I_{jt} = -15.79 + 6.23 * 0 + 56.77 * 0 + 10.61 * 1 = -5.18$$

$$E[Y_{Q4}] = \delta + \sum_{j=1}^{3} \gamma_j I_{jt} = -15.79 + 6.23 * 0 + 56.77 * 0 + 10.61 * 0 = -15.79$$

11.9 The first step is to select the appropriate model to use. For these, the statistical significance of the parameters for the constant and trend must be taken into account. For the trend model, both of these have *t*-statistics with absolute values > 4, well within the bounds of statistical significance at even the 99% level. Accordingly, the proper model to study is the last one.

For this model, the *t*-stat on γ is to the left of the 1% CV (i.e., the test statistic is more negative than the critical value so is in the rejection region)—therefore, the null hypothesis of having a unit root is rejected at the 99% confidence level. Note that if the proper model were either the constant or no-trend then the null hypothesis would not be rejected.

11.10 **a.** Given that the equation has a unit root means that the factorization of the characteristic equation must be of the form:

$$(z - 1)(z - c) = z^2 - (c + 1)z + c,$$

where $0 < |c| < 1$

The characteristic equation for the original relationship is

$$z^2 - kz - m$$

Therefore, it is required that $|m| < 1$

and

$$k = (-m + 1) \text{ or } (1 - m)$$

b. Express the equation as an AR(1) process:

$$Y_t = (-m + 1)Y_{t-1} + mY_{t-2} + \epsilon_t$$

$$Y_t - Y_{t-1} = -mY_{t-1} + mY_{t-2} + \epsilon_t$$

$$= -m(Y_{t-1} - Y_{t-2}) + \epsilon_t$$

Defining $Z_t = Y_t - Y_{t-1}$

yields

$$Z_t = -mZ_{t-1} + \epsilon_t$$

11.11 **a.** H_0: $m = 0$

H_1: $m < 0$

The counterhypothesis is different from usual practice because only the left tail is being evaluated.

b. Completing the table yields

t	y_t	y_{t-1}	Δy_t	$y_{t-1} * \Delta y_t$	y_{t-1}^2
1	−1.81	0	−1.81	0.00	0.00
2	−2.31	−1.81	−0.50	0.91	3.28
3	−1.65	−2.31	0.66	−1.52	5.34
4	−0.81	−1.65	0.84	−1.39	2.72
5	−0.31	−0.81	0.50	−0.41	0.66
6	−0.37	−0.31	−0.06	0.02	0.10
7	0.47	−0.37	0.84	−0.31	0.14
8	1.10	0.47	0.63	0.30	0.22
9	0.53	1.10	−0.57	−0.63	1.21
10	0.47	0.53	−0.06	−0.03	0.28
11	1.00	0.47	0.53	0.25	0.22
12	1.68	1.00	0.68	0.68	1.00
13	0.90	1.68	−0.78	−1.31	2.82
14	1.99	0.90	1.09	0.98	0.81
15	1.31	1.99	−0.68	−1.35	3.96
16	0.76	1.31	−0.55	−0.72	1.72
17	0.43	0.76	−0.33	−0.25	0.58
18	−1.01	0.43	−1.44	−0.62	0.18
19	0.27	−1.01	1.28	−1.29	1.02
20	−2.21	0.27	−2.48	−0.67	0.07
21	−3.21	−2.21	−1.00	2.21	4.88
22	−1.90	−3.21	1.31	−4.21	10.30
23	−0.53	−1.90	1.37	−2.60	3.61
24	−0.60	−0.53	−0.07	0.04	0.28
25	0.34	−0.60	0.94	−0.56	0.36
SUM		−5.81	0.34	−12.50	45.76

So *m* can be calculated as:

$$m = \frac{\Sigma x_i y_i}{\Sigma x_i^2} = \frac{-12.50}{45.76} = -0.273$$

Returning to calculate the errors:

t	y_t	y_{t-1}	Δy_t	$y_{t-1} * \Delta y_t$	y_{t-1}^2	ϵ_t
1	−1.81	0	−1.81	0.00	0.00	−1.81
2	−2.31	−1.81	−0.50	0.91	3.28	−0.99
3	−1.65	−2.31	0.66	−1.52	5.34	0.03
4	−0.81	−1.65	0.84	−1.39	2.72	0.39
5	−0.31	−0.81	0.50	−0.41	0.66	0.28
6	−0.37	−0.31	−0.06	0.02	0.10	−0.14
7	0.47	−0.37	0.84	−0.31	0.14	0.74
8	1.10	0.47	0.63	0.30	0.22	0.76
9	0.53	1.10	−0.57	−0.63	1.21	−0.27
10	0.47	0.53	−0.06	−0.03	0.28	0.08
11	1.00	0.47	0.53	0.25	0.22	0.66
12	1.68	1.00	0.68	0.68	1.00	0.95
13	0.90	1.68	−0.78	−1.31	2.82	−0.32
14	1.99	0.90	1.09	0.98	0.81	1.34
15	1.31	1.99	−0.68	−1.35	3.96	−0.14
16	0.76	1.31	−0.55	−0.72	1.72	−0.19
17	0.43	0.76	−0.33	−0.25	0.58	−0.12
18	−1.01	0.43	−1.44	−0.62	0.18	−1.32
19	0.27	−1.01	1.28	−1.29	1.02	1.00
20	−2.21	0.27	−2.48	−0.67	0.07	−2.41
21	−3.21	−2.21	−1.00	2.21	4.88	−1.60
22	−1.90	−3.21	1.31	−4.21	10.30	0.43
23	−0.53	−1.90	1.37	−2.60	3.61	0.85
24	−0.60	−0.53	−0.07	0.04	0.28	−0.21
25	0.34	−0.60	0.94	−0.56	0.36	0.78
SUM		−5.81	0.34	−12.50	45.76	−1.25
St Dev		1.36				0.95

Using the STDEV function in EXCEL,

The standard error of this is:

$$\sigma_m = \frac{\sigma_\epsilon}{\sqrt{n}\sigma_{y_{t-1}}} = \frac{0.95}{5 * 1.36} = 0.140$$

So the t-statistic is:

$$t - stat = \frac{m}{\sigma_m} = \frac{-0.273}{0.140} = -1.96$$

As the t-statistic is less than the 5% point on the DF distribution, the null is rejected at a 95% confidence level, but the t-stat is not sufficiently negative to enable rejection at the 99% confidence level.

12

Measuring Returns, Volatility, and Correlation

Learning Objectives

After completing this reading you should be able to:

- Calculate, distinguish, and convert between simple and continuously compounded returns.
- Define and distinguish among volatility, variance rate, and implied volatility.
- Describe how the first two moments may be insufficient to describe non-normal distributions.
- Explain how the Jarque-Bera test is used to determine whether returns are normally distributed.
- Describe the power law and its use for non-normal distributions.
- Define correlation and covariance and differentiate between correlation and dependence.
- Describe properties of correlations between normally distributed variables when using a one-factor model.

Financial asset return volatilities are not constant, and how they change can have important implications for risk management. For example, if returns are normally distributed with an annualized volatility of 10%, then a portfolio loss larger than 1% occurs on only 5% of days and the loss on 99.9% of days will be less than 2%. When annualized volatility is 100%, then 5% of days have losses larger than 10.2%, and one in four days would have losses exceeding 4.2%.

This chapter begins by examining the distributions of financial asset returns and why they are not consistent with the normal distribution. In fact, returns are both fat-tailed and often skewed. The heavy tails are, in particular, generated by time-varying volatility. Two popular measures of volatility are presented, the Black-Scholes-Merton model and the VIX Index. Both measures make use of option prices to measure the future expected volatility.

Capturing the dependence among assets in a portfolio is also a crucial step in portfolio construction. In portfolios with many assets, the distribution of the portfolio return is predominantly determined by the dependence between the assets held. If the assets are weakly related, then the gains to diversification are large, and the chance of experiencing an exceptionally large loss at the portfolio level should be small. On the other hand, if the assets are highly dependent, especially in their tails, then the probability of a large portfolio loss may be surprisingly high.

The final section reviews linear correlation, which is a common measure that plays a key role in optimizing a portfolio. It is not, however, enough to characterize the dependence between the asset returns in a portfolio. Instead, this section presents two alternatives that are correlation-like but have different properties. These alternative measures are better suited to understanding dependence when it is nonlinear. They also have some useful invariance and robustness properties.

12.1 MEASURING RETURNS

All estimators of volatility depend on returns, and there are two common methods used to construct returns. The usual definition of a return on an asset bought at time $t - 1$ and sold at time t is:

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}}, \qquad (12.1)$$

where P_t is the price of the asset in period t.

Returns computed using this formula are called simple returns and are traditionally expressed with an uppercase letter (i.e., R_t). The time scale is arbitrary and may be short (e.g., an hour or a day) or long (e.g., a quarter or a year).

The return of an asset over multiple periods is calculated using the product of the simple returns in each period:

$$1 + R_T = \prod_{t=1}^{T}(1 + R_t) \qquad (12.2)$$

There are also continuously compounded returns, also known as log returns. These are computed as the difference of the natural logarithm of the price:

$$r_t = \ln P_t - \ln P_{t-1} \qquad (12.3)$$

The log return is traditionally denoted with lower case letter (i.e., r_t).

The main advantage of log returns is that the total return over multiple periods is just the sum of the single period log returns:

$$r_T = \sum_{t=1}^{T} r_t \qquad (12.4)$$

A log return approximates the simple return. However, the accuracy of this approximation is poor when the simple return is large, and log returns are more commonly used when returns are computed over short time spans.

Converting between the simple and log return uses the relationship:

$$1 + R_t = \exp r_t \qquad (12.5)$$

Figure 12.1 shows the relationship between the simple return and the log return. In the left panel, which shows the relationship when the simple return is between −12% and 12%, the approximation error is small (i.e., less than 0.85%). The right panel expands the range to ±50%, and the approximation error is much larger (e.g., when the simple return is −40%, the log return is −51%; and when the simple return is 40%, the log return is 33.6%).

The log return is always less than the simple return. Furthermore, simple returns are also never less than −100% (i.e., a total loss). Log returns do not respect this intuitive boundary and are less than −100% whenever the simple return is less than −63%.

The simple additive structure of log returns makes them more attractive in cases where the error is small (i.e., over short periods). For example, it is common for volatility models to use log returns because they are usually computed using daily or weekly data. Despite this common practice, however, it is important to have a sense of the approximation error embedded in log returns when returns are large in magnitude.

An advantage of simple returns is that the simple return on a portfolio over a given time period is just the weighted average of the simple returns for its component assets (i.e., they are cross-sectionally additive). This does not apply for continuously compounded returns, however, since taking a log is a nonlinear transformation and thus the sum of a log is not the same as the log of a sum. Instead, the geometric mean is needed to obtain the continuously compounded return (as in equation 12.2).

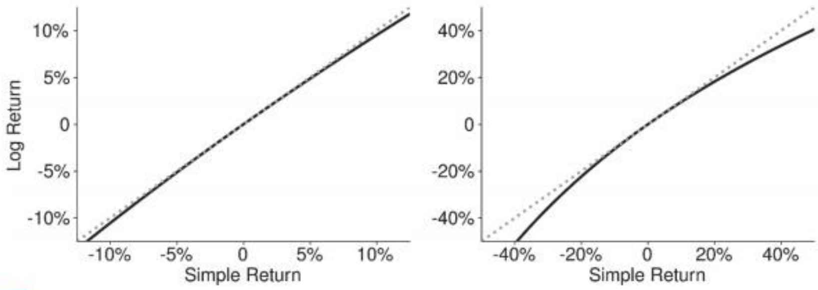


Figure 12.1 The left panel shows the relationship between the simple return and the log return when the simple return ranges between −12% and 12%. The right panel expands the range of the returns to −50% and 50%.

12.2 MEASURING VOLATILITY AND RISK

The volatility of a financial asset is usually measured by the standard deviation of its returns. A simple but useful model for the return on a financial asset is

$$r_t = \mu + \sigma e_t, \tag{12.6}$$

where e_t is a shock with mean zero and variance 1, μ is the mean of the return (i.e., $E[r_t] = \mu$), and σ^2 is the variance of the return (i.e., $V[r_t] = \sigma^2$).

The shock can also be written as:

$$\epsilon_t = \sigma e_t$$

so that ϵ_t has mean zero and variance σ^2.

The shock is assumed to be independent and identically distributed (iid) across observations. While it is not necessary to specify the distribution of e_t, it is often assumed that $e_t \overset{iid}{\sim} N(0,1)$. With this assumption, returns are also iid $N(\mu, \sigma^2)$. However, while the normal is a convenient distribution to use, the next section demonstrates that it does not provide a full description of most financial returns.

The volatility of an asset is usually measured using the standard deviation of the returns (i.e., $\sqrt{\sigma^2} = \sigma$). This measure is the volatility of the returns over the time span where the returns are measured, and so if returns are computed using daily closing prices, this measure is the daily volatility.

In a basic model, the return over multiple periods is constructed as the sum of the returns. For example, if returns are calculated daily, the weekly return is:

$$\sum_{i=1}^{5} r_{t+i} = \sum_{i=1}^{5} \mu + \sigma e_{t+i} = 5\mu + \sigma \sum_{i=1}^{5} e_{t+i}$$

The mean of the weekly return is 5μ and, because e_t is an iid sequence, the variance of the weekly return is $5\sigma^2$ and the volatility of the weekly return is $\sqrt{5}\sigma$. This is an important feature of financial returns—both the mean and the variance scale linearly in the holding period, while the volatility scales with the square-root of the holding period.[1]

This scaling law allows us to transform volatility between time scales. The common practice is to report the *annualized volatility*, which is the volatility over a year. When the volatility is measured daily, it is common to convert daily volatility to annualized volatility by scaling by $\sqrt{252}$ so that:

$$\sigma_{\text{annual}} = \sqrt{252 \times \sigma^2_{\text{daily}}} \tag{12.7}$$

The scaling factor 252 approximates the number of trading days in the US and most other developed markets. Volatilities computed over other intervals can be easily converted by computing the number of sampling intervals in a year and using this as the scaling factor.

For example, if volatility is measured with monthly returns, then the annualized volatility is constructed using a scaling factor of 12:

$$\sigma_{\text{annual}} = \sqrt{12 \times \sigma^2_{\text{monthly}}} \tag{12.8}$$

The variance, also called the variance rate, is just the square of the volatility. It scales linearly with time and so can also be easily transformed to an annual rate.

[1] The linear scaling of the variance depends crucially on the assumption that returns are not serially correlated. This assumption is plausible for the time series of returns on many liquid financial assets.

Table 12.1 **Annualized Volatilities Computed from Variance Estimates Constructed Daily, Weekly, and Monthly for the Returns on Gold, the S&P 500, and the JPY/USD Exchange Rate**

	Gold	S&P 500	¥/$
Daily	15.0%	16.0%	8.5%
Weekly	17.6%	17.1%	9.7%
Monthly	17.1%	14.5%	9.8%

Consider the volatility of the returns of three assets:

1. Gold (specifically, the price as determined at the London gold fix),
2. The return on the S&P 500 index, and
3. The percentage change in the Japanese yen/US dollar rate (JPY/USD).

Returns are computed daily, weekly (using Thursday closing prices), and monthly (using the final closing price in the month).[2] The variance of returns is estimated using the standard estimator:

$$\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}(r_t - \hat{\mu})^2$$

applied to returns constructed at each of the three frequencies, where $\hat{\mu}$ is the sample average return.

Table 12.1 shows the annualized volatilities constructed by transforming the variance estimates. The annualized volatilities are similar within each asset and the minor differences across sampling frequencies can be attributed to estimation error (because these volatilities are all estimates of the true volatility).

Implied Volatility

Implied volatility is an alternative measure that is constructed using option prices. Both put and call options have payouts that are nonlinear functions of the underlying price of an asset. For example, the payout of a European call option is defined as:

$$\max(P_T - K, 0), \qquad (12.9)$$

where P_T is the price of the asset when the option expires (T) and K is called the strike price.

This expression is a convex function of the price of the asset and so is sensitive to the variance of the return on the asset.

The most well-known expression for determining the price of an option is the Black-Scholes-Merton model. It relates the price of a call option (i.e., C_t) to the interest rate of a riskless asset (i.e., r_f), the current asset price, the strike price, the time until maturity (measured in T years), and the annual variance of the return (i.e., σ^2). All values in the model, including the call price, are observable *except* the volatility:

$$C_t = f(r_f, T, P_t, K, \sigma^2) \qquad (12.10)$$

The value of σ that equates the observed call option price with the four other parameters in the formula is known as the implied volatility. This implied volatility is, by construction, an annual value, and so it does not need to be transformed further.

The Black-Scholes-Merton option pricing model uses several simplifying assumptions that are not consistent with how markets actually operate. Specifically, the model assumes that the variance does not change over time. This is not compatible with financial data, and so the implied volatility that is calculated using the model is not always consistent across options with different strike prices and/or maturities.

The VIX Index is another measure that reflects the implied volatility on the S&P 500 over the next calendar 30 days, constructed using options with a wide range of strike prices.

The methodology of the VIX has been extended to many other assets, including other key equity indices, individual stocks, gold, crude oil, and US Treasury bonds. The most important limitation of the VIX is that it can only be computed for assets with large, liquid derivatives markets, and so it is not possible to apply the VIX methodology to most financial assets. Because the VIX Index makes uses of option prices with expiration dates in the future, it is a forward-looking measure of volatility. This differs from backward-looking volatility estimates generated using (historical) asset returns.

12.3 THE DISTRIBUTION OF FINANCIAL RETURNS

A normal distribution is symmetric and thin-tailed, and so has no skewness or excess kurtosis. However, many return series are both skewed and fat-tailed.

For example, consider the returns of gold, the S&P 500, and the JPY/USD exchange rate. The left two columns of Table 12.2 contain the skewness and kurtosis of these series when measured using daily, weekly (Thursday-to-Thursday), monthly, and quarterly returns. The monthly and quarterly returns are computed using the last trading date in the month or quarter, respectively.

All three assets have a skewness that is different from zero and a kurtosis larger than 3. In general, the skewness changes little as the horizon increases from daily to quarterly, whereas the

[2] Weekly returns use Thursday-to-Thursday prices to avoid the impact of any end-of-week effects.

Table 12.2 Estimated Skewness and Kurtosis for Three Assets Using Data between 1978 and 2017. Returns Are Computed from Prices Sampled Daily, Weekly, Monthly, and Quarterly. The Third Column Contains the Value of the Jarque-Bera Test Statistic. The Final Column Contains the p-Value for the Null Hypothesis That H_0:$S = 0$ and $\kappa = 3$, Where S Is the Skewness and κ Is the Kurtosis

		Skewness	Kurtosis	JB Statistic	p-Value
Gold	**Daily**	0.31	17.57	89571.3	0.000
	Weekly	0.93	13.08	9139.2	0.000
	Monthly	0.63	7.05	358.6	0.000
	Quarterly	1.03	6.65	116.7	0.000
S&P 500	**Daily**	−0.74	23.79	182548.7	0.000
	Weekly	−0.68	11.68	6715.3	0.000
	Monthly	−0.65	5.26	136.2	0.000
	Quarterly	−0.62	3.99	16.6	0.000
JPY/USD	**Daily**	−0.38	6.95	6766.7	0.000
	Weekly	−0.61	7.13	1610.9	0.000
	Monthly	−0.21	4.09	27.4	0.000
	Quarterly	−0.20	3.14	1.2	0.557

kurtosis declines substantially. The returns on the S&P 500 and the returns on the JPY/USD exchange rate are both negatively skewed, while the returns on gold are positively skewed. This difference arises because equities and gold tend to move in different directions when markets are stressed—equities decline markedly in stress periods, while gold serves as a flight-to-safety asset and so moves in the opposite direction.

The Jarque-Bera test statistic is used to formally test whether the sample skewness and kurtosis are compatible with an assumption that the returns are normally distributed. The null hypothesis in the Jarque-Bera test is H_0:$S = 0$ and $\kappa = 3$, where S is the skewness and κ is the kurtosis of the return distribution. These are the population values for normal random variables. The alternative hypothesis is H_1:$S \neq 0$ or $\kappa \neq 3$. The test statistic is:

$$JB = (T - 1)\left(\frac{\hat{S}^2}{6} + \frac{(\hat{\kappa} - 3)^2}{24}\right), \tag{12.11}$$

where T is the sample size.

When returns are normally distributed, the skewness is asymptotically normally distributed with a variance of 6, so that $\hat{S}^2/6$ has a χ^2_1 distribution. Meanwhile, the kurtosis is asymptotically normally distributed with mean 3 and variance of 24, and so $(\hat{\kappa} - 3)^2/24$ also has a χ^2_1 distribution. These two statistics are asymptotically independent (uncorrelated), and so $JB \sim \chi^2_2$.

Test statistics that have χ^2 distributions should be small when the null hypothesis is true and large when the null is unlikely to be correct. The critical values of a χ^2_2 are 5.99 for a test size of 5% and 9.21 for a test size of 1%. When the test statistic is above these values, the null that the data are normally distributed is rejected.

The JB statistic is reported for the three assets in Table 12.2. It is rejected in all cases except for the quarterly return on the JPY/USD rate. The final column reports the p-value of the test statistic, which is $1 - F_{\chi^2_2}(JB)$, where $F_{\chi^2_2}$ is the CDF of a χ^2_2 random variable.

While the test rejects for equities even at the quarterly frequency, the low-frequency returns are closer to a normal than the daily or weekly returns on the S&P 500. This is a common feature of many financial return series—returns computed over longer periods appear to be better approximated by a normal distribution.

Power Laws

An alternative method to understand the non-normality of financial returns is to study the tails. Normal random variables have thin tails so that the probability of a return larger than $k\sigma$ declines rapidly as k increases, whereas many other distributions have tails that decline less quickly for large deviations.

The most important class of these have *power law* tails, so that the probability of seeing a realization larger than a given value of x is:

$$P(X > x) = kx^{-\alpha},$$

where k and α are constants.

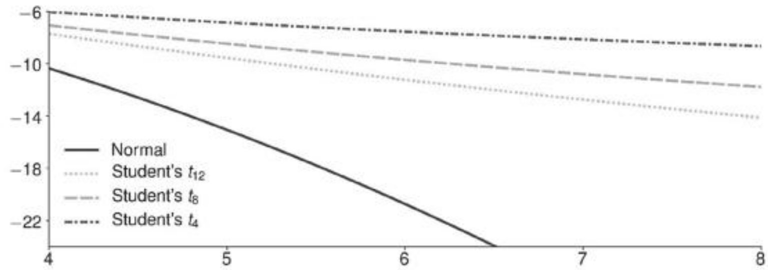


Figure 12.2 **Plot of the log probability that a random variable $X > x$, where X has a mean of zero and unit standard deviation. Each curve represents $\ln Pr(X > x) = \ln(1 - F_X(x))$, which measures the chance of appearing in the extreme upper tail for $x \in (4, 8)$.**

The Student's t is an example of a widely used distribution with a power law tail.

The simplest method to compare tail behavior is to examine the natural log of the tail probability. Figure 12.2 plots the natural log of the probability that a random variable with mean zero and unit variance appears in its tail [i.e., $\ln \Pr(X > x)$]. Four distributions are shown: a normal and three parameterizations of a Student's t. For example, the log probability of a random variable being > 6 is around -20 for the normal distribution but -7 for a t-distribution with 4 degrees of freedom. To obtain the actual probabilities from these log probabilities, use the exponential (i.e., e^{-20} and e^{-7} for the normal and t-distribution, respectively).

The normal curve is quadratic in x, and its tail quickly diverges from the tails of the other three distributions. The Student's t tails are linear in x, which reflects the slow decay of the probability in the tail. This slow decline is precisely why these distributions are fat-tailed and produce many more observations that are far from the mean when compared to a normal.

12.4 CORRELATION VERSUS DEPENDENCE

The dependence between assets plays a key role in portfolio diversification and tail risk. Recall that two random variables, X and Y, are independent if their joint density is equal to the product of their marginal densities:

$$f_{X,Y}(x, y) = f_X(x)f_Y(y)$$

Any random variables that are not independent are dependent. Financial assets are highly dependent and exhibit both linear and nonlinear dependence. The linear correlation estimator (also known as Pearson's correlation) measures linear dependence. Previous chapters have shown that correlation (i.e., ρ) and the regression slope (i.e., β) are intimately related, and the regression slope is zero if and only if the correlation is zero. Regression explains the sense in which correlation measures linear dependence.

In the regression:

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

if Y and X are standardized to have unit variance (i.e., $\sigma_X^2 = \sigma_Y^2 = 1$), then the regression slope *is* the correlation.

In contrast, nonlinear dependence takes many forms and cannot be summarized by a single statistic. For example, many asset returns have common heteroskedasticity (i.e., the volatility across assets is simultaneously high or low). However, linear correlation does not capture this type of dependence between assets.

Alternative Measures of Correlation

Linear correlation is insufficient to capture dependence when assets have nonlinear dependence. Researchers often use two alternative dependence measures: rank correlation (also known as Spearman's correlation) and Kendal's τ (tau). These statistics are correlation-like: both are scale invariant, have values that always lie between -1 and 1, are zero when the returns are independent, and are positive (negative) when there is in increasing (decreasing) relationship between the random variables.

Rank Correlation

Rank correlation is the linear correlation estimator applied to the *ranks* of the observations. Suppose that n observations of two random variables, X and Y, are obtained. Let $Rank_X$ and $Rank_Y$ be the rank of X and Y, respectively. The rank operator assigns 1 to the smallest value of each, 2 to the second smallest, and so

on until the largest is given rank n.[3] The rank correlation estimator is defined as:

$$\hat{\rho}_s = \frac{\widehat{\text{Cov}}[Rank_X, Rank_Y]}{\sqrt{\hat{\text{V}}[Rank_X]}\sqrt{\hat{\text{V}}[Rank_Y]}} \quad (12.12)$$

When all ranks are distinct, the estimator can be equivalently expressed as a function of the difference in the pairwise ranks:

$$\hat{\rho}_s = 1 - \frac{6\sum_{i=1}^{n}(Rank_{X_i} - Rank_{Y_i})^2}{n(n^2-1)}, \quad (12.13)$$

where $Rank_{X_i} - Rank_{Y_i}$ measures the difference in the ranks of the two elements of the same observation (i.e., X_i and Y_i). When highly ranked values of X are paired with highly ranked values of Y, then the difference is small (i.e., $(R_{X_i} - R_{Y_i})^2 \approx 0$) and the correlation is near 1. When X and Y have strong negative dependence (so that the largest values of X couple with the smallest values of Y), then the difference is large and the rank correlation is close to -1.

Note that rank correlation depends on the strength of the linear relationship between the *ranks* of X and Y, not the linear relationship between the random variables themselves.

When variables have a linear relationship, rank and linear correlation are usually similar in magnitude. The rank correlation estimator is less efficient than the linear correlation estimator and is commonly used as an additional robustness check.

Significant differences in the two correlations indicate an important nonlinear relationship. Rank correlation has two distinct advantages over linear correlation:

1. First, it is robust to outliers because only the ranks, not the values of X and Y, are used.
2. Second, it is invariant with respect to any monotonic increasing transformation of X_i and Y_i (e.g., $X_i \rightarrow f(X_i)$, where $f(X_i) > f(X_j)$ when $X_i > X_j$). Linear correlation is only invariant with respect to increasing linear transformations (e.g., $X_i \rightarrow a + bX_i$, where $b > 0$). This invariance with respect to a wide class of transformations makes rank correlation particularly useful when examining the relationship between primary and derivative asset returns.

Figure 12.3 illustrates the difference between correlation and rank correlation. The top-left panel plots simulated observations from the data-generating process (DGP) $Y_i = X_i + \epsilon_i$, where X_i and ϵ_i are independent standard normal random variables. The top-right panel plots simulated data from the DGP $Y_i = \exp(2X_i) + \epsilon_i$ using the same realizations of X_i and ϵ_i. Both plots include the true conditional mean $E[Y|X]$ and the fitted regression line $\hat{\alpha} + \hat{\beta}X$. These align closely in the left panel and only differ due to estimation error. In the right panel, however, the linear relationship is a poor approximation to the conditional expectation.

The bottom panels plot the ranked values $Rank_X$ against $Rank_Y$. Both DGPs are monotonic and so the ranks have linear relationships.

Kendall's τ

Kendall's τ measures the relative frequency of *concordant* and *discordant* pairs.

Two random variables, (X_i, Y_i) and (X_j, Y_j) are concordant if $X_i > X_j$ and $Y_i > Y_j$ or if $X_i < X_j$ and $Y_i < Y_j$. In other words, concordant pairs agree about the relative position of X and Y. Pairs that disagree about the order are discordant, and pairs with ties ($X_i = X_j$ or $Y_i = Y_j$) are neither. Random variables with a strong positive dependence have many concordant pairs, whereas variables with a strong negative relationship have mostly discordant pairs.

Kendal's τ is defined as:

$$\hat{\tau} = \frac{n_c - n_d}{n(n-1)/2} = \frac{n_c}{n_c + n_d + n_t} - \frac{n_d}{n_c + n_d + n_t}, \quad (12.14)$$

where n_c is the count of the concordant pairs, n_d is the count of the discordant pairs, and n_t is the count of the ties.[4]

Kendall's τ is defined as the difference between the probability of concordance and the probability of discordance. When all pairs are concordant, then $\hat{\tau} = 1$. If all pairs are discordant, $\hat{\tau}$ is -1. Other patterns produce values between these two extremes. Note that concordance only depends on the pairwise order, not the magnitude of data, and so monotonic increasing transformations also have no effect on Kendall's τ.

Figure 12.4 plots the relationship between linear (Pearson) correlation, rank (Spearman) correlation, and Kendall's τ for a bivariate normal.

The rank correlation is virtually identical to the linear correlation for normal random variables. Kendall's τ is increasing in ρ, although the relationship is not linear. Kendall's τ produces values much closer to zero than the linear correlation for most of the range of

[3] When observation values are not unique then the ranks cannot be uniquely assigned, and so each observation in a tied group is assigned the average rank of the observations in the group. For example, in the data set {1, 2, 3, 3, 3, 3, 4, 4, 5}, the value 3 repeats four times. The ranks of the four values in the group of 3s are 3, 4, 5, and 6, so that the rank assigned to each 3 is $(3 + 4 + 5 + 6)/4 = 4.5$. The two repetitions of 4 are assigned rank $(7 + 8)/2 = 7.5$, and so that the ranks of the values in the data set are {1, 2, 4.5, 4.5, 4.5, 4.5, 7.5, 7.5, 9}.

[4] This estimator may be far from 1 or -1 if the data have many ties. There are improved estimators that modify the denominator to account for ties that have better properties when this is the case. Ties are not an important concern in most applications to financial assets returns, and so it is not usually necessary to use an alternative.

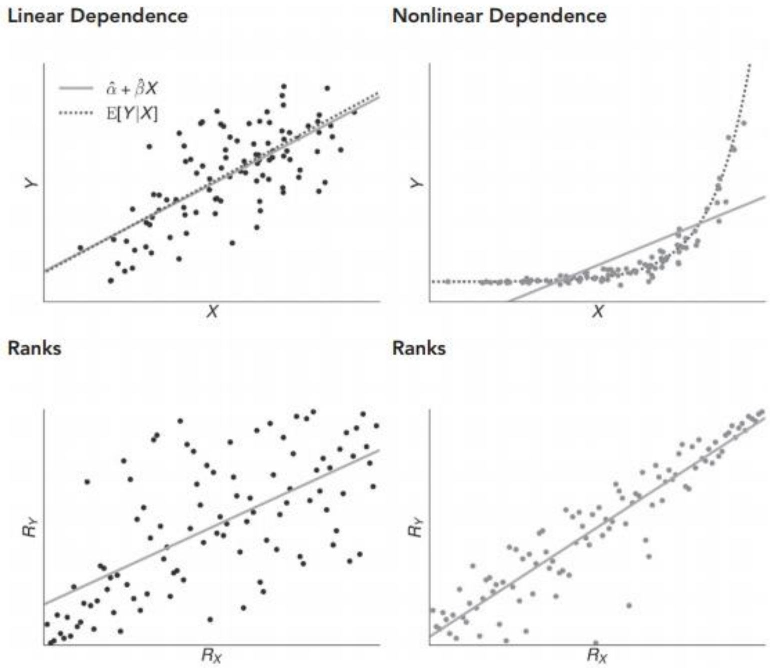


Figure 12.3 The left panels plot data that are linearly dependent. The top-left panel plots the raw data along with the conditional expectation and the estimated regression line. The bottom-left panel shows the ranks of the data and the fitted regression line on the ranks. The right panels show data with a nonlinear relationship. The top-right panel shows the data, the conditional expectation, and the linear approximation. The bottom-right panel plots the ranks of the data along with the fitted regression line.

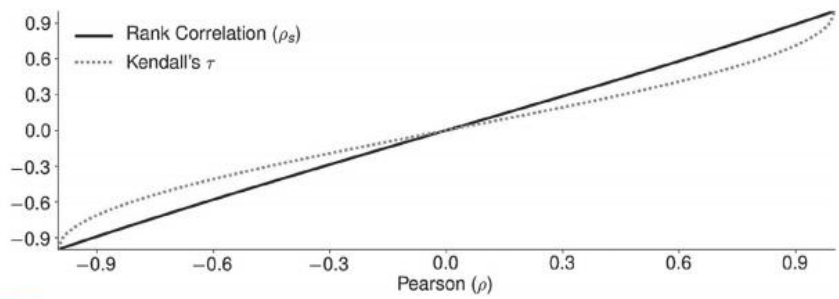


Figure 12.4 Relationship between Pearson's correlation (ρ), Spearman's rank correlation (ρ_s), and Kendall's τ for a bivariate normal distribution.

ρ. Estimates of τ with the same sign as the sample correlation, but that are smaller in magnitude, do not provide evidence of nonlinearity.

Table 12.3 contains the population values for the linear correlation, rank correlation, and Kendall's τ using the two DGPs depicted in Figure 12.3. The linear and rank correlations are nearly identical when the DGP is linear and Kendall's τ is predictably lower here. When the DGP is nonlinear, both the rank correlation and Kendall's τ are larger than the linear correlation. This pattern in the two alternative measures of correlation indicates that the dependence in the data has important nonlinearities.

Table 12.3 **Population Values of Linear (Pearson) Correlation, Rank (Spearman) Correlation, and Kendall's τ for the Linear and Nonlinear DGPs in Figure 12.3**

	Linear DGP	Nonlinear DGP
Linear Correlation	0.707	0.291
Rank Correlation	0.690	0.837
Kendall's τ	0.500	0.667

STRUCTURED CORRELATION MATRICES

A correlation matrix is interpretable as a covariance matrix where all random variables have unit variance. The fact that any linear combination of random variables must have a non-negative variance imposes restrictions on the values in a correlation matrix of three or more random variables.

For example, suppose that a trivariate normal random variable with components X_1, X_2, and X_3 has mean zero and covariance (and correlation) matrix:

$$\Sigma = R = \begin{bmatrix} 1 & -0.9 & -0.9 \\ -0.9 & 1 & -0.9 \\ -0.9 & -0.9 & 1 \end{bmatrix}.$$

The variance of an average of the three components:

$$V[\bar{X}] = \frac{1}{9}(3 \times 1 + 2 \times -0.9 + 2 \times -0.9 + 2 \times -0.9) = -0.267$$

This value is negative, and so Σ *cannot* be a valid covariance matrix. This requirement—that any weighted average must have a positive variance—is known as positive definiteness.

Practitioners commonly impose structure on correlation matrices to ensure that they are positive definite. Two structured correlations are commonly used.

1. The first type sets all correlations equal to the same value (i.e., $\rho_{ij} = \bar{\rho}$ for all pairs i and j), a model known as 'equicorrelation'.
2. The second type uses a factor structure that assumes correlations are due to exposure to a common factor. This type of structured correlation can mimic the correlation of assets related through the CAPM, for example, so that the correlation between any two entries can be written as $\rho_{ij} = \gamma_i\gamma_j$, where both γ_1 and γ_2 are between -1 and 1.

12.5 SUMMARY

This chapter began with a discussion of simple returns and log returns. These two measures broadly agree when returns are small but substantially disagree if they are large in magnitude. In practice, log returns are used when prices are sampled frequently, because the approximation error is unimportant. Over longer horizons, the simple return provides a more accurate measure of the performance of an investment.

Both the mean and the variance of asset returns grow linearly in time, so that the h-period mean and variance are h multiplied by the first-period mean and variance. The standard deviation, which is the square root of the variance, grows with $\sqrt{h}$.

Implied volatility is an alternative measure of risk that can be constructed for assets that have liquid options markets. Implied volatility can be estimated using the Black-Scholes-Merton option pricing model or the VIX Index, which reflects the implied volatility on the S&P 500 over the next 30 calendar days.

While an assumption of normality is convenient, it is not consistent with most observed financial asset return data. In fact, most returns are both skewed and heavy-tailed. These two features are most prominent when measuring returns using high-frequency data (e.g., daily), whereas the normal is a better approximation over long horizons (i.e., one quarter or longer).

Deviations from normality can be examined in a couple of ways. The Jarque-Bera test examines the skewness and kurtosis of a data series to determine whether it is consistent with an assumption of normality. Meanwhile, the tail index is a measure of the tail decay that examines the rate at which the log-density decays. Small values of the tail index indicate that the returns on an asset are heavy-tailed.

The chapter concluded with an examination of alternative measures of dependence. While linear correlation is an important measure, it is only a measure of linear dependence. Two alternative measures, rank correlation and Kendall's τ, are more suited to measure nonlinear dependence. These two measures are correlation-like in the sense that they are between -1 and 1. Both also have two desirable properties when compared to linear correlation: They are invariant to monotonic increasing transformations (both linear and nonlinear) and are robust to outliers.

The following questions are intended to help candidates understand the material. They are *not* actual FRM exam questions.

QUESTIONS

Short Concept Questions

12.1 If returns are fat-tailed, and we instead assume that they are normally distributed, is the probability of a large loss over- or under-estimated?

12.2 Is linear correlation sensitive to a large outlier pair? What about rank correlation and Kendall's τ?

12.3 Variances are usually transformed into volatilities by taking the square root. This changes the units from squared returns to returns. Why isn't the same transformation used on covariances?

Practice Questions

12.4 On Black Monday in October 1987, the Dow Jones Industrial Average fell from the previous close of 2,246.74 to 1,738.74. What was the simple return on this day? What was the log return?

12.5 If the annualized volatility on the S&P 500 is 20%, what is the volatility over one day, one week, and one month?

12.6 If the mean return on the S&P 500 is 9% per year, what is the mean return over one day, one week, and one month?

12.7 If an asset has zero skewness, what is the maximum kurtosis it can have to not reject normality with a sample size of 100 using a 5% test? What if the sample size is 2,500?

12.8 If the skewness of an asset's return is −0.2 and its kurtosis is 4, what is the value of a Jarque-Bera test statistic when $T = 100$? What if $T = 1{,}000$?

12.9 Calculate the simple and log returns for the following data:

Time	Price
0	100
1	98.90
2	98.68
3	99.21
4	98.16
5	98.07
6	97.14
7	95.70
8	96.57
9	97.65
10	96.77

12.10 The implied volatility for an ATM money option is reported at 20%, annualized. Based upon this, what would be the daily implied volatility?

12.11 The following data are collected for four distributions:

Dataset	Skew	Kurtosis	T
A	0.85	3.00	50
B	0.85	3.00	51
C	0.35	3.35	125
D	0.35	3.35	250

Which of these datasets are likely (at the 95% confidence level) not to be drawn from a normal distribution?

12.12 Calculate the rank correlations for the following data:

i	X	Y
1	0.22	2.73
2	1.41	6.63
3	−0.30	−2.19
4	−0.59	−6.51
5	−3.08	−0.99
6	1.08	2.63
7	−0.45	−3.40
8	0.40	5.10
9	−0.75	−5.14
10	0.24	1.14

12.13 Find *Kendall's* τ for the following data:

j	X	V
1	3.12	2.58
2	−1.26	−0.05
3	2.08	−0.72
4	−0.28	−0.52
5	−1.96	−0.40

The following questions are intended to help candidates understand the material. They are *not* actual FRM exam questions.

ANSWERS

Short Concept Questions

12.1 We will underestimate the probability of a large loss because the normal has thin tails. For example, the probability of a 4σ loss from a normal is 0.003% (3 in 100,000). The probability of a 4σ loss from a standardized Student's t_4 is 0.24% or 80 times more likely.

12.2 Yes. Linear correlation is sensitive in the sense that moving a single point in X_1, X_2 space far from the true linear relationship can severely affect the correlation. Effectively, a single outlier has an unbounded effect on the covariance, and ultimately the linear correlation. Rank correlation and Kendall's τ are both less sensitive to a single outlier because the effect is bounded and is averaged out by the other observations. If n is large, then a single outlier will have an immaterial impact on the estimates.

12.3 Covariance may have either sign, and so the square root transformation is not reliable if the sign is negative. Transformation to β or correlation is preferred when examining the magnitude of a covariance.

Solved Problems

12.4 The simple return is $100 \times \frac{1738.74 - 2246.74}{2246.74} =$ -22.6%. The log return is $100 \times \ln 1738.74 -$ $- \ln 2246.74 = 25.6\%$. The difference between the two is increasing in the magnitude.

12.5 The scale factors for the variance are 252, 52, and 12, respectively. The volatility scales with the square root of the scale factor, and so $\frac{20\%}{\sqrt{252}} = 1.26\%$, $\frac{20\%}{\sqrt{52}} = 2.77\%$, and $\frac{20\%}{\sqrt{12}} = 5.77\%$.

12.6 The mean scales linearly with the scale factor, and so $\frac{9\%}{252} = 0.036\%$, $\frac{9\%}{52} = 0.17\%$, and $\frac{9\%}{12} = 0.75\%$.

12.7 The Jarque-Bera has χ^2_2 distribution, and the critical value for a test with a size of 5% is 5.99. The Jarque-Bera statistic is $JB = (T-1)\left(\frac{\hat{S}^2}{6} + \frac{(\hat{\kappa}-3)^2}{24}\right)$ so that when the skewness $\hat{S} = 0$, the test statistic is $(T-1)\frac{(\hat{\kappa}-3)^2}{24}$. In order to not reject the null, we need $(T-1)\frac{(\hat{\kappa}-3)^2}{24} \leq 5.99$, and so $(\hat{\kappa}-3)^2 \leq 24 \times \frac{5.99}{T-1}$ and $\hat{\kappa} \leq 3 + \sqrt{24 \times \frac{5.99}{T-1}}$. When $T = 100$, this value is 4.20. When $T = 2{,}500$ this value is 3.24. This shows that the JB test statistic is sensitive to even mild excess kurtosis.

12.8 Using the formula in the previous problem, the value of the JB is 4.785 when $T = 100$ and 48.3 when $T = 1000$.

12.9 Simple returns use the formula:

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}}$$

Log returns use the formula:

$$r_t = \ln P_t - \ln P_{t-1} = \ln\left(\frac{P_t}{P_{t-1}}\right)$$

The two sets of returns are given in the final columns of the following table:

Time	Price	Simple	Log
0	100		
1	98.90	−1.10%	−1.11%
2	98.68	−0.22%	−0.22%
3	99.21	0.54%	0.54%
4	98.16	−1.06%	−1.06%
5	98.07	−0.09%	−0.09%
6	97.14	−0.95%	−0.95%
7	95.70	−1.48%	−1.49%
8	96.57	0.91%	0.90%
9	97.65	1.12%	1.11%
10	96.77	−0.90%	−0.91%

12.10 Using the equation:

$$\sigma_{annual} = \sqrt{252 \times \sigma^2_{daily}}$$

$$0.2 = \sqrt{252 \times \sigma^2_{daily}}$$

$$\sigma_{daily} = \frac{0.2}{\sqrt{252}} = 1.26\%$$

12.11 The appropriate test statistic is the Jarque-Bera formula

$$JB = (T-1)\left(\frac{\hat{S}^2}{6} + \frac{(\hat{\kappa}-3)^2}{24}\right)$$

This will be distributed as a χ^2_2, and for a 5% test size the critical value is 5.99. The test statistic values for the four datasets are given in the final column of the following table.

Dataset	Skew	Kurtosis	T	JB
A	0.85	3.00	50	5.90
B	0.85	3.00	51	6.02
C	0.35	3.35	125	3.16
D	0.35	3.35	250	6.35

So, as the JB figure is greater than 5.99 for datasets B & D, their respective null hypotheses (normally distributed) are rejected.

12.12 Following the procedure outlined in the text (and example), the first step is to complete the table by calculating the ranks of each observation (R_{Xi} and R_{Yi}) within its individual series and then calculate the difference between the ranks for each observation *i*:

i	X	Y	R_{Xi}	R_{Yi}	$R_{Xi}-R_{Xi}$
1	0.22	2.73	6	8	−2
2	1.41	6.63	10	10	0
3	−0.30	−2.19	5	4	1
4	−0.59	−6.51	3	1	2
5	−3.08	−0.99	1	5	−4
6	1.08	2.63	9	7	2
7	−0.45	−3.40	4	3	1
8	0.40	5.10	8	9	−1
9	−0.75	−5.14	2	2	0
10	0.24	1.14	7	6	1
			Sum Squares		32

$$\hat{\rho}_s = 1 - \frac{6\sum_{i=1}^{n}(R_{Xi}-R_{Yi})^2}{n(n^2-1)} = 1 - \frac{6 * 32}{10(100-1)} = 0.81$$

12.13 The first step is to rank the two series separately:

i	X	Y	R_{Xi}	R_{Yi}
1	3.12	2.58	5	5
2	−1.26	−0.05	2	4
3	2.08	−0.72	4	1
4	−0.28	−0.52	3	2
5	−1.96	−0.40	1	3

The next step is to see which pairs are concordant, which are discordant, and which are neither. This leads to the pair-by-pair comparison:

	1	2	3	4	5
1					
2	c				
3	c	d			
4	c	d	d		
5	c	c	d	d	

For the first observation, because both the rank of X and of Y are the maximum, every other observation will automatically be concordant with respect to this one.

By contrast, looking at the second and third observations, the X rank increases while the Y rank decreases. Therefore, this pair is discordant.

$$\hat{\tau} = \frac{n_c - n_d}{n(n-1)/2} = \frac{5-5}{10} = 0$$

13

Simulation and Bootstrapping

Learning Objectives

After completing this reading you should be able to:

- Describe the basic steps to conduct a Monte Carlo simulation.
- Describe ways to reduce Monte Carlo sampling error.
- Explain the use of antithetic and control variates in reducing Monte Carlo sampling error.
- Describe the bootstrapping method and its advantage over Monte Carlo simulation.
- Describe pseudo-random number generation.
- Describe situations where the bootstrapping method is ineffective.
- Describe the disadvantages of the simulation approach to financial problem solving.

Simulation is an important practical tool in modern risk management with a wide variety of applications. Examples of these applications include computing the expected payoff of an option, measuring the downside risk in a portfolio, and assessing estimator accuracy.

Monte Carlo simulation—often called a Monte Carlo experiment or just a Monte Carlo—is a simple approach to approximate the expected value of a random variable using numerical methods. A Monte Carlo generates random draws from an assumed data generating process (DGP). It then applies the function (or functions) to these simulated values to generate a realization from the unknown distribution of the transformed random variables. This process is repeated, and the statistic of interest (e.g., the expected price of a call option) is approximated using the simulated values. Importantly, repeating the simulation improves the accuracy of the approximation, and the number of replications can be chosen to achieve any required level of precision.

Another important application of simulation methods is bootstrapping. The bootstrap gets its name from a seemingly impossible feat: "to pull oneself up by one's bootstraps." Bootstrapping uses observed data to simulate from the unknown distribution generating the observed data. This is done by combining observed data with simulated values to draw repeated samples from the original data. This creates a set of new samples, each of which is closely related to, but different from, the observed data.

Monte Carlo simulation and bootstrapping are closely related. In both cases, the goal is to compute the expected value of an (often complex) function. Both methods use computer-generated values (i.e., simulated data) to numerically approximate this expected value.

The fundamental difference between simulation and bootstrapping is the source of the simulated data. When using simulation, the user specifies a complete DGP that is used to produce the simulated data. In an application of the bootstrap, the observed data are used directly to generate the simulated data set without specifying a complete DGP.

13.1 SIMULATING RANDOM VARIABLES

Conducting simulation experiments requires generating random values from an assumed distribution. However, numbers generated by computers are not actually random and are usually referred to as pseudo-random numbers.

Pseudo-random numbers are generated by complex but deterministic functions that produce values that are difficult to predict, and so appear to be random. The functions that produce pseudo-random values are known as pseudo-random number generators (PRNGs) and are initialized with what is called a seed value. It is important to note that repeatedly using the same seed value would produce an identical set of random values every time the PNRG is run.

The reproducibility of outputs from PRNGs has two important applications.

1. It allows results to be replicated across multiple experiments, because the same sequence of random values can always be generated by using the same seed value. This feature is important when exploring alternative models or estimators, because it allows the alternatives to be estimated on the same simulated data. It also allows simulation-based results to be reproduced later, which is important for regulatory compliance.
2. Setting the seed allows the same set of random numbers to be generated on multiple computers. This feature is widely used when examining large portfolios containing many financial instruments (i.e., thousands or more) that are all driven by a common set of fundamental factors. Distributed simulations that assess the risk in the portfolio must use the same simulated values of the factors when studying the joint behavior of the instruments held in the portfolio. Initializing the PRNGs with a common seed in a cluster environment ensures that the realizations of the common factors are identical.

13.2 APPROXIMATING MOMENTS

Monte Carlo simulation is frequently used to estimate population moments or functions.

Suppose $X \sim F_X$, where F_X is a known distribution. In simple distributions, many moments (e.g., the mean or variance) are analytically tractable. For example, if $X \sim N(\mu,\sigma^2)$ then $E[X] = \mu$ and $V[X] = \sigma^2$.

However, analytical expressions for moments are not available for several relevant cases, (e.g., when random variables are constructed from complex dynamic models or when computing the expectation of a complicated nonlinear function of a random variable). Simulation provides a simple method to approximate the analytically intractable expectation.

Suppose X is a random variable that can be simulated (e.g., a normal) and g is a function that can be evaluated at realizations of X. Simulation constructs many independent copies of $g(X)$ by simulating draws from X (i.e., x_i) and then computing[1] $g_i = g(x_i)$.

[1] In many applications of simulations, X is a vector with n elements. Computing the expected value of a scalar-valued function that depends on a simulated vector uses $g_i = g(x_{1i}, x_{2i}, \ldots, x_{ni})$ in place of $g_i = g(x_i)$.

SIMULATING RANDOM VALUES FROM A SPECIFIC DISTRIBUTION

Random number generators simulate standard uniforms—Uniform(0,1). In most applications, it is necessary to transform the uniform samples into draws from another distribution.

For example, consider a simulation of data coming from a heavy-tailed probability distribution, such as a generalized Student's *t* with six degrees of freedom. Recall that a Student's *t* random variable is defined as:

$$Z/\sqrt{W/\nu},$$

where Z is a standard normal and W is an independent χ^2_ν random variable with ν degrees of freedom.

The definition of a Student's *t* cannot be directly applied because computers only generate data from uniform distributions. However, Chapter 3 showed that if a random variable X has CDF F_X, then $U = F_X(X)$ is a uniform random variable. This relationship implies that applying the inverse CDF to a uniform random variable U produces a random variable distributed according to F_X. Formally, if F_X^{-1} is the inverse CDF function of X, then $F_X^{-1}(U) \sim F_X$.

The top panel of Figure 13.1 shows how realizations of four iid uniform random variables (i.e., u_1, u_2, u_3, and u_4) are mapped through the inverse CDF (i.e., F_X^{-1}) of X to simulate four iid draws from a generalized Student's *t* with six degrees of freedom.

The bottom panel illustrates how each transformed value (i.e., x_i) is the realization that has a CDF value of u_i. This method can also be applied to discrete random variables because the CDF of a discrete random variable is a step function.

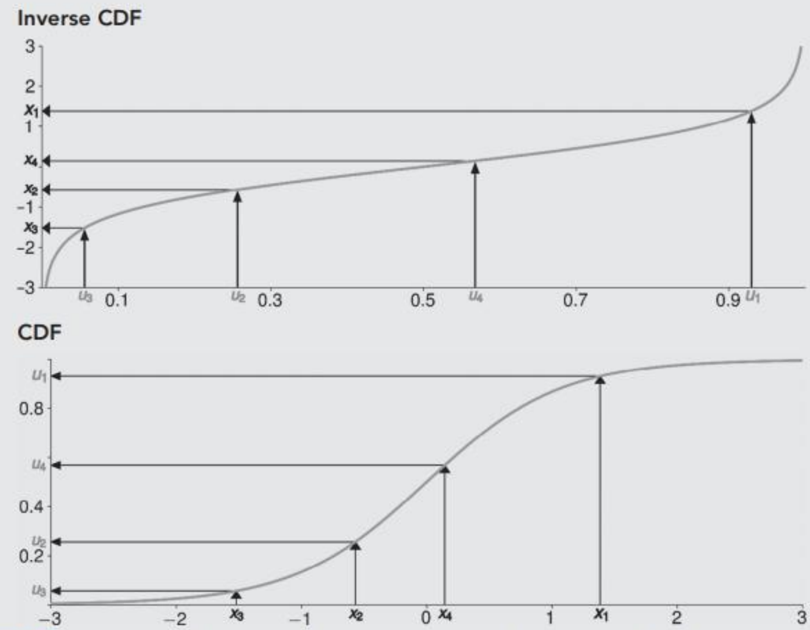


Figure 13.1 The top panel illustrates random values generated from Student's *t* distribution using uniform values, marked on the x-axis. The uniform values are mapped through the inverse CDF of the Student's *t* to generate the random variates marked on the y-axis. The bottom panel show how the generated values map back to the uniform originally used to produce each value through the CDF.

This is repeated b times to produce a set of iid draws from the unknown distribution of $g(X)$, where b is used to denote the number of replications in a simulation study and to distinguish the number of replications from the size of a single simulated sample that may have n (or T) observations. The simulated values g_i are realizations from the distribution of $g(X)$ and so can be used to approximate the moment or other quantity of interest. So in this example we are generating b sets of random draws, but each of these contains only one data point. More generally, in other applications we could generate b samples where each is a series containing n data points.

For example, to estimate the mean of $g(X)$, the unknown population value is approximated using an average. Simulated draws are iid by construction, and b draws are used to approximate the expected value as:

$$\hat{E}[g(X)] = \frac{1}{b}\sum_{i=1}^{b} g(X_i) \qquad (13.1)$$

The approximated expectation is an average of b iid random variables, and so the Law of Large Numbers (LLN) applies:

$$\lim_{b\to\infty} \hat{E}[g(X)] = E[g(X)] \qquad (13.2)$$

Furthermore, the Central Limit Theorem (CLT) also applies to the average. The variance of the simulated expectation is estimated by:

$$V[\hat{E}[g(X)]] = \sigma_g^2/b,$$

where $\sigma_g^2 = V[g(X)]$.

This variance is estimated using:

$$\hat{\sigma}_g^2 = \frac{1}{b}\sum_{i=1}^{b}(g(X_i) - E[\widehat{g(X)}])^2, \qquad (13.3)$$

which is the standard variance estimator for iid data.

The standard error of the simulated expectation (i.e., $\hat{\sigma}_g/\sqrt{b}$) measures the accuracy of the approximation, which allows b to be chosen to achieve any desired level of accuracy.

Simulations are applicable to a wide range of problems, and the set of simulated draws $\{g_1, g_2, \ldots, g_b\}$ can be used to compute other quantities of interest. For example, the α-quantile of $g(X)$ can be estimated using the empirical quantile of the simulated values. This is particularly important in risk management when estimating the lower quantiles—$\alpha \in \{0.1\%, 1\%, 5\%, 10\%\}$—of a portfolio's return distribution. The empirical α-quantile is estimated by sorting the b draws from smallest to largest and then selecting the value in position $b\alpha$ of the sorted set.

Simulation studies are also used to assess the finite sample properties of parameter estimators. Most results presented in this book rely on asymptotic results based on the LLN and the CLT. However, only n observations are availble to estimate parameters in practice, and the approximation made using a CLT may not be adequate when the sample size is small.

CONDUCTING A SIMULATION EXPERIMENT

1. Generate data $x_i = [x_{1i}, x_{2i}, \ldots, x_{ni}]$ according to the assumed DGP.
2. Calculate the function or statistic of interest, $g_i = g(x_i)$.
3. Repeat steps 1 and 2 to produce b replications.
4. Estimate the quantity of interest from $\{g_1, g_2, \ldots, g_b\}$.
5. Assess the accuracy by computing the standard error. If the accuracy is too low, increase b until the required accuracy is achieved.

Monte Carlo experiments allow the finite sample distribution of an estimator to be tabulated and compared to its asymptotic distribution derived from the CLT. The finite sample distribution is also widely used to examine the effect of using an estimate instead of the true value parameter when making a decision. Assessing the uncertainty in an estimated parameter aids in understanding the impact of alternative decision rules (e.g., whether to hedge a risk fully or partially) under realistic conditions where the true parameter is unknown. Consider the finite-sample distribution of a particular model parameter $\hat{\theta}$. First, a simulated sample of random variables $x = [x_1, x_2, \ldots, x_n]$ is generated from the assumed DGP. These simulated values are then used to estimate $\hat{\theta}$.

This process—simulating a new data set and then estimating the model parameters—is repeated b times. This produces a set of b samples $\{\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_b\}$ from the finite-sample distribution of the estimator of θ. These values can be used to assess any property (e.g., the bias, variance, or quantiles) of the finite-sample distribution of $\hat{\theta}$.

For example, the bias, defined as:

$$\text{Bias}(\theta) = E[\hat{\theta}] - \theta$$

is estimated in a finite sample of size b using:

$$\widehat{\text{Bias}}(\theta) = \frac{1}{b}\sum_{i=1}^{b}(\hat{\theta}_i - \theta) \qquad (13.4)$$

The Mean of a Normal

Consider the problem of approximating the mean of a random variable that comes from a normal probability distribution $X \sim N(\mu, \sigma^2)$. By construction, the mean of X is μ, so it is not necessary to use simulation to approximate this value. It is also

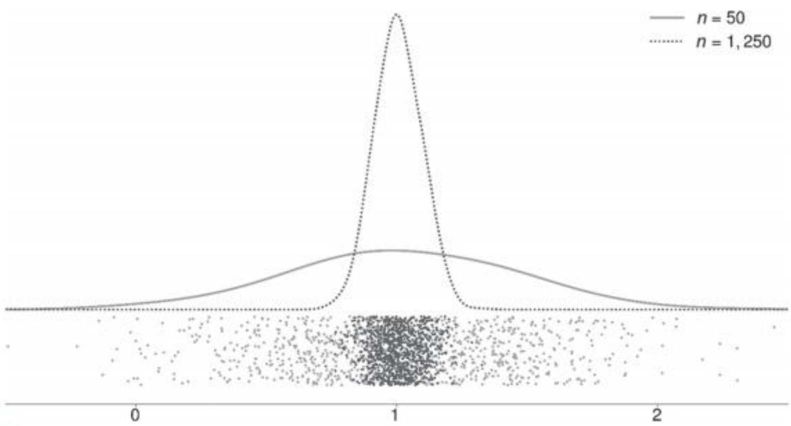


Figure 13.2 Distribution of the estimated expected value of a normal with $\mu = 1$ and $\sigma^2 = 9$ for simulated sample sizes of 50 and 1,250.

known that $E[\bar{X}] = \mu$ and $V[\bar{X}] = \sigma^2/n$. Approximating the mean in a Monte Carlo simulation requires generating b (where $b = n$) draws from a normal with mean μ and variance σ^2.

Figure 13.2 contains a density plot of $\hat{E}[X] = \bar{X}$, where $\mu = 1$ and $\sigma^2 = 3$ for a fixed number of replications ($b = 1{,}000$), each with sample sizes $n = 50$ and $n = 1{,}250$.

In both cases, the distribution is centered on the true value 1. However, realizations when $n = 50$ are highly dispersed and many are far from 1. When $n = 1{,}250$, the density is much more concentrated and virtually all realizations are within ± 0.2 of 1.

Table 13.1 contains the average estimates of μ and the standard error of the expected value estimate. The standard error decreases as b increases, although only slowly. Increasing n by a factor of 25 (i.e., from 50 to 1,250) reduces the standard error by a factor of five. This is the expected reduction because the

Table 13.1 Average Estimated Expected Value and Error of the Expected Value as a Function of the sample size *n*.

n	Mean of Sample Means	Standard Deviation of Sample Means
10	0.985	0.917
50	1.017	0.428
250	1.001	0.185
1250	1.005	0.087

sampling error of a simulated expectation is proportional to $1/\sqrt{n}$. This rate of convergence holds for any moment approximated using simulation and is a consequence of the CLT for iid data. Furthermore, this convergence rate applies even to approximations of the expected value of complex functions because the n data points are iid.

Approximating the Price of a Call Option

As an example, consider the simulation of the price of a European call option. A European call pays

$$\max(0, S_T - K),$$

where S_T is the stock price when the option expires at time T and K is the strike price.

Figure 13.3 plots the payoff of the call options as a function of the price of the underlying asset (S) on the expiry date (T). The payoff is a nonlinear function of the underlying stock price at expiration, so computing the expected payoff requires a model for the stock price.

In this case, it is assumed that the log of the stock price is normally distributed. The time T log stock price (i.e., s_T) is equal to the sum of the initial stock price s_0, a mean, and a normally distributed error:

$$s_T = s_0 + T(r_f - \sigma^2/2) + \sqrt{T}x_i \quad (13.5)$$

where x_i is a simulated value from a $N(0, \sigma^2)$, r_f is the risk-free rate, and σ^2 is the variance of the stock return. Both the risk-free rate and the variance are measured using annualized values, and T is defined so that one year corresponds to $T = 1$.

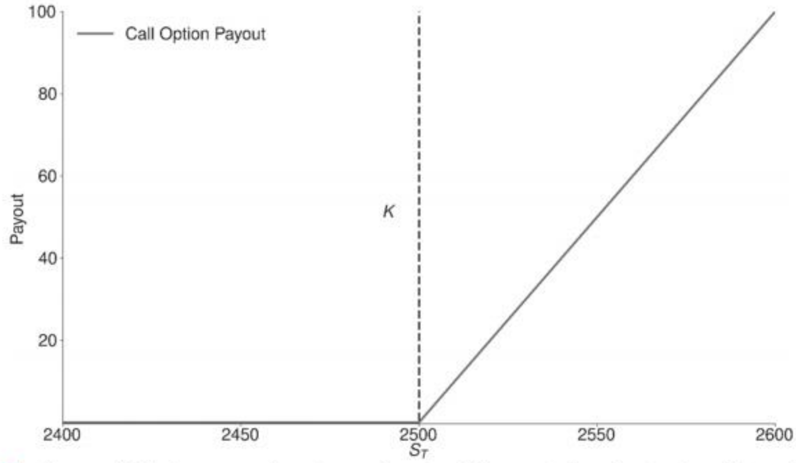


Figure 13.3 The payoff of a European call option as a function of the terminal stock price S_T and the strike price K.

The final stock price is $S_T = \exp(s_T)$, and the present value of the option's payoff is

$$C = \exp(-r_f T)\max(S_T - K, 0) \quad (13.6)$$

Simulating the price of an option requires an estimate of the stock price volatility. The estimated volatility from the past 50 years of the S&P 500 returns is $\sigma = 16.4\%$. The risk-free rate r_f is assumed to be 2%, the option expires in two years, the initial price is $S_0 = 2{,}500$, and the strike price is $K = 2{,}500$.

The expected price of a call option is estimated by the average of the simulated values:

$$\widehat{E[C]} = \overline{C} = \frac{1}{b}\sum_{i=1}^{b} C_i,$$

where C_i are simulated values of the payoff of the call option.

Using Equation 13.5, the simulated stock prices are

$$S_{Ti} = \exp(s_0 + T(r_f - \sigma^2/2) + \sqrt{T}x_i)$$

Using the notation of the previous section, the function $g(x_i)$ is defined

$$g(x_i) = \exp(-r_f T)\max(\exp(s_0 + T(r_f - \sigma^2/2) + \sqrt{T}x_i) - K, 0)$$

The number of replications varies from $b = 50$ to 3,276,800 in constant multiples of 4. The left column of Table 13.2 shows the expected price of a call option from a single approximation, and the center column reports the estimated standard error of the approximation. The standard error is computed as:

$$\text{s.e.}(\widehat{E[C]}) = \sqrt{\hat{\sigma}_g^2/b}, \quad (13.7)$$

where $\hat{\sigma}_g^2 = \frac{1}{b}\sum_i (C_i - \overline{C})^2$ is the variance estimator applied to the simulated call prices.

A two-sided 95% confidence interval for the price can be constructed using the average simulated price and its standard error:

$$[\widehat{E[C]} - 1.96 \times \text{s.e.}(\widehat{E[C]}), \widehat{E[C]} + 1.96 \times \text{s.e.}(\widehat{E[C]})]$$

Table 13.2 Estimated Present Value of the Payoff of a European Call Option with Two Years to Maturity. The Left Column Contains the Approximation of the Price Using b Replications. The Center Column Reports the Estimated Standard Error of the Approximation. The Right Column Reports the Ratio of the Standard Error in the Row to the Standard Error in the Row Above

b	Price	Std. Err.	Std. Err. Ratio
50	301.50	71.23	–
200	283.63	30.19	2.36
800	276.21	15.39	1.96
3,200	283.93	7.48	2.06
12,800	275.62	3.64	2.06
51,200	278.49	1.85	1.97
204,800	278.64	0.93	1.99
819,200	278.85	0.46	2.01
3,276,800	279.15	0.23	2.00

The standard error is large relative to the expected option price except for the largest values of b. For example, when $b = 12{,}800$, the 95% confidence interval covers a range of USD 14 ($=2 \times 1.96 \times 3.64$), which is 5% of the value of the call option.

The right column of Table 13.2 contains the ratio of the standard error of the estimate with the next smallest replication size. The standard error declines in the expected manner as b increases because the standard error is $\sigma_g/\sqrt{b}$, where σ_g is the standard deviation of the simulated call option prices.

Improving the Accuracy of Simulations

The standard method of estimating an expected value using simulation relies on the LLN. The standard error of the estimated expected value is proportional to $1/\sqrt{b}$ and, because the simulations are independent, it only depends on the variance of the simulated values.

One way to increase the accuracy of estimation from Monte Carlo simulation is to increase the number of replications. If the simulation is complex, however, this can be computationally slow. Other ways to improve Monte Carlo sampling accuracy for a given number of replications include using either antithetic variates or control variates (or both).

Recall that the variance of a sum of random variables is the sum of the variances plus twice the covariances between each distinct pair of random variables. When the simulations are independent, these covariances are all zero. However, if the covariance between the simulated values is negative, then the variance of the sum is less than the sum of the variances. This is the idea behind antithetic variables, which are random values that are constructed to generate negative correlation within the values used in the simulation.

Control variates are an alternative that add values that are mean zero and correlated to the simulation. Because the control variate is mean zero, its addition does not bias the approximation. The correlation between the control variate and the function of interest allows an optimal combination of the control variate and the original simulation value to be constructed to reduce the variance of the approximation.

Antithetic variables and control variates are complementary, and they can be used simultaneously to reduce the approximation error in a Monte Carlo simulation.

Antithetic Variates

Antithetic variates add a second set of random variables that are constructed to have a negative correlation with the iid variables used in the simulation. They are generated in pairs using a single uniform value.

Recall that if U_1 is a uniform random variable, then:

$$F_X^{-1}(U_1) \sim F_X$$

Antithetic variables can be generated using:

$$U_2 = 1 - U_1,$$

where U_2 is also a uniform random variable and, so:

$$F_X^{-1}(U_2) \sim F_X$$

By construction, the correlation between U_1 and U_2 is negative, and mapping these values through the inverse CDF generates F_X distributed random variables that are negatively correlated.

For example, suppose that F is a standard normal distribution. Recall that the standard normal is symmetric so that $F_X(u_1) = -F_X(1 - u_1)$ (e.g., if $u_1 = 0.25$, then $x_1 = F^{-1}(u_1) = -0.67$ and $x_2 = F^{-1}(u_2) = F^{-1}(1 - u_1) = 0.67$). These values are perfectly negatively correlated.

Using antithetic random variables in simulations is virtually identical to running a standard simulation. The only difference occurs when generating the values used in the simulations. These values are generated in pairs: $(U_1, 1 - U_1)$, $(U_2, 1 - U_2), \ldots, (U_{b/2}, 1 - U_{b/2})$. These uniforms are then transformed to have the required distribution using the inverse CDF.

Note that only $b/2$ uniforms are required to produce b simulated values because the random variables are generated in pairs. Antithetic pairs produced using this algorithm are only guaranteed to reduce the simulation error if the function $g(X)$ is monotonic (i.e., either always increasing or decreasing) in x. Monotonicity ensures that the negative correlation between x_i and $-x_i$ will translate to $g(x_i)$ and $g(-x_i)$.

The improvement in accuracy when using antithetic variables occurs due to the correlation between the paired random variables. Essentially, the technique works by helping to fill up the space in the probability distribution more evenly and may be computationally more efficient than simply increasing the number of replications. In the usual case with b iid draws, the standard error of the simulated expectation is $\sigma_g/\sqrt{b}$. Since the antithetic variables are correlated, the standard error of the simulated expectation becomes

$$\frac{\sigma_g\sqrt{1+\rho}}{\sqrt{b}} \tag{13.8}$$

Thus, the standard error is reduced if $\rho < 0$.

Control Variates

Control variates are an alternative method to reduce simulation error. The standard simulation method uses the approximation:

$$\hat{E}[g(X)] = \frac{1}{b}\sum_{i=1}^{b} g(x_i)$$

This approximation is consistent because each random variable $g(X_i)$ can be decomposed as

$$E[g(X_i)] + \eta_i,$$

where η_i is a mean zero error.

A control variate is another random variable $h(X_i)$ that has mean 0 (i.e., $E[h(X_i)] = 0$) but is correlated with the error η_i (i.e., $Corr[h(X_i), \eta_i] \neq 0$).

A good control variate should have two properties.

1. First, it should be inexpensive to construct from x_i. If the control variate is slow to compute, then larger variance reductions—holding the computational cost fixed—may be achieved by increasing the number of simulations b rather than constructing the control variates.
2. Second, a control variate should have a high correlation with $g(X)$. The optimal combination parameter β that minimizes the approximation error is estimated using the regression:

$$g(x_i) = \alpha + \beta h(x_i) + \nu_i \quad (13.9)$$

Application: Valuing an Option

Reconsider the example of simulating the price of a European call option.

The left-center columns of Table 13.3 show the prices simulated using antithetic random variables, along with their standard errors. Because the random values used in the simulation are normally distributed, the antithetic value is the negative of the original value. In other words, the values used in the simulation are $x_1, -x_1, x_2, -x_2, \ldots, x_{b/2}, -x_{b/2}$.

The value of the call option is a nonlinear function of these values, so the correlation of the simulated call option prices is not the same as the correlation of the simulated values of X. The correlation ρ between the paired call option prices that use x_i and $-x_i$ (i.e., the correlation between $g(x_i)$ and $g(-x_i)$, where $g(X)$ is the call option value function) is determined to be -43%. The standard error of the simulated value is reduced by approximately 25% ($1 - \sqrt{1 + \rho}$). In this example, using antithetic variables is equivalent to performing 78% more simulations using only independent draws and so has almost doubled the number of replications at a relatively low computational cost.

The right-center columns of Table 13.3 show the effect of using a control variate (in this case, the value of a European put option). The application of control variates involves employing a variable similar to that used in the simulation but whose properties are already known. The simulation is conducted on both the variable under study and on the known variable, with the same sets of random number draws being employed in both cases. The idea is that if the control and simulation variables are closely related, then the effects of sampling error for the problem under study and the known problem will be similar, and hence can be reduced by calibrating the Monte Carlo results using the analytic results. A put option has the payoff function $\max(0, K - S_T)$, so it is in the money when the final stock price is below the strike price K. Here, the control variate is constructed by subtracting the closed-form Black-Scholes-Merton put price from the simulated put price:

$$\tilde{P}_i = \exp(-rT)\max(0, K - S_T) - P^{BS}(r, S_0, K, \sigma, T)$$

Table 13.3 **Estimated Present Value of the Payoff of a European Call Option with Two Years to Maturity. The Left Columns Use b iid Simulations. The Results in the Left-Center Columns Use Antithetic Variables. The Results in the Right-Center Columns Add a Control Variate to the Simulation. The Results in the Right-Most Columns Use Both Antithetic Variables and a Control Variate**

	Standard		Antithetic		Control Var.		Both	
b	**Price**	**Std. Err.**	**Price**	**Std. Err.**	**Price**	**Std. Err.**	**Price**	**Std. Err.**
50	301.50	71.23	388.58	65.46	326.31	64.10	433.40	58.38
200	283.63	30.19	286.49	23.61	287.51	26.49	289.33	21.08
800	276.21	15.39	279.95	11.43	276.32	13.71	280.07	10.21
3,200	283.93	7.48	282.06	5.54	282.95	6.61	284.01	4.89
12,800	275.62	3.64	279.12	2.75	277.72	3.21	279.68	2.43
51,200	278.49	1.85	278.99	1.39	279.46	1.64	279.14	1.24
204,800	278.64	0.93	278.39	0.70	278.53	0.83	278.18	0.62
819,200	278.85	0.46	278.06	0.35	278.55	0.41	277.79	0.31
3,276,800	279.15	0.23	278.45	0.17	278.89	0.21	278.36	0.15

The simulated price of the put estimates the analytical price, and so the difference is mean zero (i.e., $E[\bar{P}] = 0$). The call option price is then approximated by:

$$\hat{E}[C] = \frac{1}{b}\sum_{i=1}^{b} C_i + \hat{\beta}\left(\frac{1}{b}\sum_{i=1}^{b} \bar{P}_i\right),$$

where $\hat{\beta}$ is the OLS estimate computed by regressing C_i on $\bar{P}_i$.

Note that the variance reduction is similar in magnitude to the reduction achieved using antithetic variables, because the correlation between the simulated call and put option prices, (i.e., $\text{Corr}[C_i, \bar{P}_i]$) is determined to be −44%.

The right-most columns of Table 14.3 use both antithetic variables and the control variate to reduce the standard error by 33% when compared to a simulation using b iid draws. The combined effect is equivalent to using 125% more iid replications.

Limitation of Simulations

Monte Carlo simulation is a straightforward method to approximate moments or to understand estimators' behavior. The biggest challenge when using simulation to approximate moments is the specification of the DGP. If the DGP does not adequately describe the observed data, then the approximation of the moment may be unreliable. The misspecification in the DGP can occur due to many factors, including the choice of distributions, the specification of the dynamics used to generate the sample, or the use of imprecise parameter estimates to simulate the data.

The other important consideration is the computational cost. Running a single simulation for a simple DGP takes little time on a modern computer. However, running many simulation experiments (e.g., to understand the sensitivity of an option price to model parameters) can be very time-consuming. This cost limits the ability of an analyst to explore the specification used for the DGP. Therefore, analytical expressions should be preferred when available, and simulation should be used when they are not.

13.3 BOOTSTRAPPING

Bootstrapping is an alternative method to generate simulated data. However, while both simulation and bootstrapping make use of observed data, there are some important differences.

Simulation uses observed data for calibration only—in other words, to estimate key model parameters (e.g., the mean and standard deviation of the S&P 500). These parameters are then combined with an assumption about the distribution of the standardized returns to complete the DGP.

In contrast, bootstrapping directly uses the observed data to simulate a sample that has similar characteristics. Specifically, it does this without directly modeling the observed data or making specific assumptions about their distribution. In this sense, the only thing that is being simulated in a bootstrap is the choice of observations to select when constructing the samples.

The key to understanding bootstrapping lies in one simple fact: The unknown distribution being sampled from is the same one that produced the observed data. While this distribution is unknown, a distribution that resembles (in a probabilistic sense) the distribution that produced the observed data can be sampled from by using the same observed data.

There are two classes of bootstraps used in risk management. The first is known as the iid bootstrap. It is extremely simple to implement because the samples are created by drawing with replacement from the observed data.

Suppose that a simulation sample size of m is needed from a data set with a total of n observations. The iid bootstrap generates observation indices by randomly sampling with replacement from the values $\{1, 2, \ldots, n\}$. These random indices are then used to select the observed data that are included in the simulated sample. The resampled data are commonly described as a bootstrap sample to distinguish them from the original data.

For example, suppose that five observations are required from a sample of 20 data points $\{x_1, \ldots, x_{20}\}$. Table 13.4 contains a set of five bootstrap simulations, each with five samples.

The first simulation uses observations $\{x_3, x_6, x_2, x_5, x_{10}\}$, and the second simulation uses $\{x_6, x_3, x_2, x_{18}, x_{15}\}$. These two simulated samples partially overlap. This is expected because the iid bootstrap samples with replacement. The fifth simulation sample uses $\{x_8, x_{12}, x_8, x_2, x_5\}$, so it repeats the same observation twice (i.e., x_8). This is also expected when generating many bootstrap samples and the probability[2] of an index repeating is nearly 1.

The bootstrap is otherwise identical to Monte Carlo simulation, where the bootstrap samples are used in place of the simulated samples. The expected value of a function is approximated by:

$$\hat{E}[g(X)] = \frac{1}{b}\sum_{j=1}^{b} g(x_{1,j}^{BS}, x_{2,j}^{BS}, \ldots, x_{m,j}^{BS}), \tag{13.10}$$

where $x_{i,j}^{BS}$ is observation i from bootstrap sample j and a total of b bootstrap replications are used.

[2] In an iid bootstrap with b replications where each bootstrap sample has m observations from an original sample size of n, the probability of at least one bootstrap sample with a repeated index is:

$$1 - \left(\frac{n!}{n^m(n-m)!}\right)^b,$$

where ! is the factorial operator. In the example in the text with five bootstrap samples, each with a size of five drawn from a sample size of 20, this value is 93.4%.

Table 13.4 Simulated Indices Used to Produce Five iid Bootstrap Samples, Each with Size Five from an Original Sample with 20 Observations

Locationn	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
1	3	6	10	14	8
2	6	3	17	11	12
3	2	2	19	18	8
4	5	18	10	19	2
5	10	15	12	7	5

GENERATING A SAMPLE USING THE iid BOOTSTRAP

1. Generate a random set of m integers, $\{i_1, i_2, \ldots, i_m\}$, from $\{1, 2, \ldots, n\}$ with replacement.
2. Construct the bootstrap sample as $\{x_{i_1}, x_{i_2}, \ldots, x_{i_m}\}$.

The iid bootstrap is generally applicable when observations are independent across time. In some situations, this is a reasonable assumption. For example, if sampling from the S&P 500 in 2005 or 2017, both calm periods with low volatility, the distribution of daily returns is similar throughout the entire year and so the iid bootstrap is a reasonable choice.

However, it is often the case that financial data are dependent over time and thus a more sophisticated bootstrapping method is required to allow for this dependence. Otherwise, the properties of the bootstrapped samples will differ from those of the actual data. The simplest of these methods is known as the circular block bootstrap (CBB). It only differs in one important aspect—instead of sampling a single observation with replacement, the CBB samples blocks of size q with replacement.

For example, suppose that 100 observations are available and they are sampled blocks of size $q = 10$. 100 blocks are constructed, starting with $\{x_1, x_2, \ldots, x_{10}\}$, $\{x_2, x_3, \ldots, x_{11}\}, \ldots,$ $\{x_{91}, x_{92}, \ldots, x_{100}\}$, $\{x_{92}, x_{93}, \ldots, x_{100}, x_1\}, \ldots, \{x_{100}, x_1, \ldots x_9\}$. Note that while the first 91 blocks use ten consecutive observations, the final nine blocks all wrap around (as if observation one occurs immediately after observation 100).

The CBB is used to produce bootstrap samples by sampling blocks, with replacement, until the required bootstrap sample size is produced. If the number of observations sampled is larger than the required sample size, the final observations in the final block are dropped. For example, if the block size is ten and the desired sample size if 75, then eight blocks of ten are selected and the final five observation from the final block are dropped.

GENERATING A SAMPLE USING THE CBB

1. Select the block size q.
2. Select the first block index i from $\{1, 2, \ldots, n\}$ and append $\{x_i, x_{i+1}, \ldots x_{i+q}\}$ to the bootstrap sample where indices larger than n wrap around.
3. While the bootstrap sample has fewer than m elements, repeat step 2.
4. If the bootstrap sample has more than m elements, drop values from the end of the bootstrap sample until the sample size is m.

The final detail required to use the CBB is the choice of the block size q. This value should be large enough to capture the dependence in the data, although not so large as to leave too few blocks. The rule-of-thumb is to use a block size equal to the square root of the sample size (i.e., $\sqrt{n}$).

Bootstrapping Stock Market Returns

Estimating market risk is an important application of bootstrapping in financial markets. Here we are interested in estimating the p-Value-at-Risk (p-VaR):

$$\underset{VaR}{\operatorname{argmin}} \Pr(L > VaR) = 1 - p, \tag{13.11}$$

where L is the loss of the portfolio over a specified period (e.g., one month) and $1 - p$ is the probability that the loss occurs.

When the loss is measured as a percentage of the portfolio value, then the p-VaR is equivalently defined as a quantile of the return distribution.

As an example, consider a one-year VaR for the S&P 500. Computing the VaR requires simulating many copies of one year of data (i.e., 252 days) and then finding the empirical quantile of the simulated annual returns.

The data set used spans from 1969 until 2018. Both the iid and the CBBs are used to generate bootstrap samples. The data set

is large (i.e., 13,000 observations) and so the block size for CBB is set to 126 observations (i.e., six months of daily data).

Bootstrapped annual log returns are simulated using 252 daily log returns:

$$r_j^A = \sum_{i=1}^{252} r_{ij}^{BS},$$

where the r_{ij}^{BS} are the bootstrapped log returns for replication j.

The VaR is then computed by sorting the b bootstrapped annual returns from lowest to highest and then selecting the value at position $(1 - p)b$, which is the empirical $1 - p$ quantile of the annual returns.

The number of bootstrap replications is $b = 100{,}000$. A large number of replications is needed to improve the accuracy of the 0.1% quantile because only one in 1,000 replications falls into this extreme left tail and therefore the estimate effectively relies only on a very small proportion of the total number of points generated.

As shown in Table 13.5, the CBB produces larger VaR estimates for small values of $1 - p$ and smaller VaR estimates for large values when compared to the iid bootstrap. This difference occurs because financial asset returns are dependent. Specifically, financial assets show evidence of volatility clustering and so go through periods where volatility is consistently higher (e.g., 2002 or 2008) or lower (e.g., 2005 or 2017) than its long-run average.

The iid bootstrap samples at random from the entire set of observations and in doing so produces a simulated sample that is too "balanced." A typical sample from the iid bootstrap has five returns from each of the 50 years in the sample. In contrast, the CBB uses exactly two blocks to simulate an annual return. When the CBB selects a block, the volatility—high, normal, or low—is preserved throughout the block.

The inclusion of long periods with elevated volatility explains why the VaR estimates for $p = 99.9\%$ or 99% from the CBB are larger than the estimates from the iid bootstrap. This same feature also explains the lower VaR estimates when $p = 95\%$ or less. The smaller VaR estimates are the result of the CBB selecting blocks with volatility below average, which produce simulated losses smaller than those produced by the "balanced" (but not empirically realistic) returns constructed using the iid bootstrap.

Figure 13.4 shows the left tail of the simulated return distribution, focusing on the bottom 10% of the distribution of bootstrapped annual returns.

Table 13.5 **VaR Estimates (Quantiles of the Annual Return Distribution) Using the iid and Circular Block Bootstraps**

Quantile	Value-at-Risk iid	Value-at-Risk (CBB)
.1%	44.1%	53.5%
1%	31.0%	35.6%
5%	19.0%	18.2%
10%	12.9%	10.9%

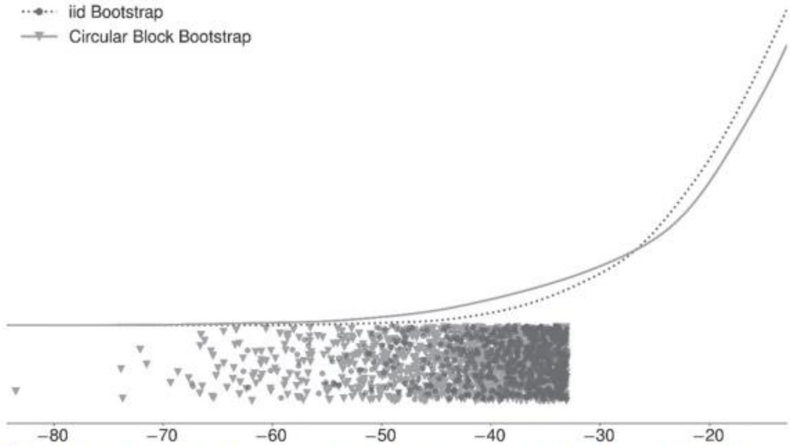


Figure 13.4 **Left tail of the simulated annual return distribution using the iid and circular block bootstraps. The tail shows 10% of the return density. The points below the density show the lowest 1% of the simulated annual returns produced by the two bootstrap methods.**

The tails of the densities cross near the 2% quantile, and the CBB produces a density of annual returns with more probability in the tail. This crossing is another consequence of the excess balance in the iid bootstrap samples, and the distribution constructed from the iid bootstrap is less heavy-tailed than the distribution constructed from the CBB bootstrap. The points below the curves show the smallest 1% of actual simulated annual returns used to estimate the VaR. Only 36% of these extreme observations are generated by the iid bootstrap.

Limitations of Bootstrapping

While bootstrapping is a useful statistical technique, it has its limitations. There are two specific issues that arise when using a bootstrap.

- First, bootstrapping uses the entire data set to generate a simulated sample. When the past and the present are similar, this is a desirable feature. However, if the current state of the financial market is different from its normal state, then the bootstrap may not be reliable. For example, many assets experienced record levels of volatility during the financial crisis of 2008. Using a bootstrap to estimate the VaR in October 2008 produces an optimistic view of risk because the bootstrap estimate uses data from periods where the volatility was much lower.
- The second limitation arises due to structural changes in markets so that the present is significantly different from the past. Consider bootstrapping the interest rate on short-term US government debt. This rate was near zero from late 2008 until 2015 due to both market conditions and interventions by the Federal Reserve. Bootstrapping the historical data cannot produce samples with short-term interest rates that replicate this period because rates had never previously been near zero for an extended period.

13.4 COMPARING SIMULATION AND BOOTSTRAPPING

Monte Carlo simulation and bootstrapping are both essential tools in finance and risk management. They are closely related and only differ in the method used to generate simulated data.

A Monte Carlo simulation uses a full statistical model that includes an assumption about the distribution of the shocks. The need for a complete model is a weakness of simulation, and poor models produce inaccurate results even when the number of replications is large.

The bootstrap method avoids the specification of a model and instead makes the key assumption that the present resembles the past. The validity of this assumption is essential when using a bootstrap. Applying a bootstrap requires understanding the dependence in the observed data, and it is essential that the bootstrap method used replicates the actual dependence observed in the data. If the bootstrap does not reproduce the dependence, then statistics computed from bootstrapped samples cannot capture the sampling variation in the observed data.

Both Monte Carlo simulation and bootstrapping suffer from the "Black Swan" problem—simulations generated using either method resemble the historical data. Bootstrapping is especially sensitive to this issue, and a bootstrap sample cannot generate data that did not occur in the sample. A good statistical model, on the other hand, should allow the possibility of future losses that are larger than those that have been realized in the past. In practice, DGPs used in simulation studies are calibrated to historical data and so are usually limited in their ability to generate simulation samples substantially different from what has been already observed.

13.5 SUMMARY

This chapter introduces simulation as a method to approximate moments or other quantities that depend on the expected value of a random variable (e.g., an option price) when analytical expressions are not available.

Two closely related methods are covered: Monte Carlo simulation and bootstrapping. Both methods use random samples to simulate data: A simulation uses a complete specification for the DGP, whereas the bootstrap method randomly resamples from the observed data. However, the key steps to approximate a moment are the same for the two methods.

1. Construct a simulated or bootstrap sample.
2. Compute the function or estimator using the sample created in step 1.
3. Repeat steps 1 and 2 to construct b replications.
4. Estimate the quantity of interest using the b values produced in step 2.

Both simulation and the bootstrap depend on user choices. In a simulation, the entire DGP is specified—this DGP plays a key role in the validity of the simulation experiment. When using a bootstrap, the resampled data must mimic the dependence structure of the data, and the past must be a reliable guide for the future.

The following questions are intended to help candidates understand the material. They are *not* actual FRM exam questions.

QUESTIONS

Short Concept Questions

13.1 What is the main difference between a simulation and the bootstrap?

13.2 How is the seed of a pseudo-random number generator useful?

13.3 What are antithetic variables and control variates?

13.4 Why do antithetic variables and control variates improve the accuracy of a simulation?

13.5 How are bootstrap samples generated using the iid bootstrap?

13.6 How are bootstrap samples generated using the circular block bootstrap (CBB)?

13.7 When is the CBB needed?

13.8 What are the key limitations of the bootstrap?

Practice Questions

13.9 Suppose that you are interested in approximating the expected value of an option. Based on an initial sample of 100 replications, you estimate that the fair value of the option is USD 47 using the mean of these 100 replications. You also note that the standard deviation of these 100 replications is USD 12.30. How many simulations would you need to run in order to obtain a 95% confidence interval that is less than 1% of the fair value of the option? How many would you need to run to get within 0.1%?

13.10 Plot the variance reduction when using antithetic random variables as a function of the correlation between pairs of observations.

13.11 Plot the variance reduction when using control variates as a function of the correlation between the control variate and the random variable used in the simulation.

13.12 A random draw from an N(1,5) distribution is 1.2. If this was generated using the inverse of the cumulative distribution approach, what is the original draw from the U(0,1) distribution?

13.13 Given the data below:

	N(0,1)
1	−0.65
2	−0.12
3	0.92
4	1.28
5	0.63
6	−1.98
7	0.22
8	0.4
9	0.86
10	1.74

a. Calculate the option payoffs using the equation Max(x − 0.5,0), which is equivalent to calculating the value of a call option on the expiry date if the underlying value is $S_T = x$ and the strike price K is 0.5.

b. What are the antithetic counterparts for each scenario and their associated values?

c. Assuming that the sample correlation holds for the entire population, what is the change in expected sample standard error attainable by using antithetic sampling?

d. How does the answer to part c change if the payoff function is: Max($|x - 0.5|$, 0)?

The following questions are intended to help candidates understand the material. They are *not* actual FRM exam questions.

ANSWERS

Short Concept Questions

13.1 A simulation uses artificial data by drawing from a specific distribution to simulate the shocks and other quantities in a model error. A simulation requires a fully specified model that provides the distribution of all random quantities. A bootstrap uses random sampling from the observed data to produce a new dataset that has similar characteristics. In a bootstrap, the random samples are used to generate indices when selecting the data to include in the bootstrap sample.

13.2 The seed is a value that sets the internal state of a pseudo-random number generator. Setting the seed allows the same sequence of pseudo-random numbers to be generated, which allow simulations to be reproduced exactly.

13.3 Antithetic variables are simulated variables that have been constructed to have a negative correlation with a previously sampled set of simulated values. Control variates are mean zero values that are correlated with the quantity whose expectation is being computed via simulation. The control variates are then combined with the simulated value using an optimal weight to reduce the simulation variance. Control variates should be cheap to produce.

13.4 Both introduce some negative correlation in the simulation. Antithetic variables are selected to have negative correlation across simulations. Because the variance of the sum depends on the covariance, and the covariance is negative, the variance of the sum is less than the case if the variables were independent. Control variates add a mean-zero term to a simulation that has a negative correlation with the simulated values. The variance reduction works in the same manner because the variance of the sum depends on the sum of the variances plus the covariances. If the covariances are negative, the Monte Carlo sampling variance is reduced.

13.5 A set of n integers are randomly generated with replacement from $1, 2, \ldots, n$. These are used to select the indices of the original data to produce the bootstrapped sample.

13.6 The CBB uses samples from the original data in blocks of size m. Each block is chosen by selecting the index for the first observation in the block. A total of $\lceil \frac{n}{m} \rceil$ blocks are selected, where n is the sample size where $\lceil \cdot \rceil$ selects the smallest integer larger than its input. If the number of observations selected is larger than m due to rounding, only the first n are kept. If a block starts in the last $m - 1$ observations, the block wraps around to the start of the sample.

13.7 The CBB is generally needed when the data being bootstrapped have time series dependence—in other words, a block bootstrap is required when the statistic being bootstrapped is sensitive to the serial correlation in the data. For example, when bootstrapping the mean, the CBB is required if the data are serially correlated. If bootstrapping the variance, the CBB is required if the squares are serially correlated—that is, if the data has conditional heteroskedasticity.

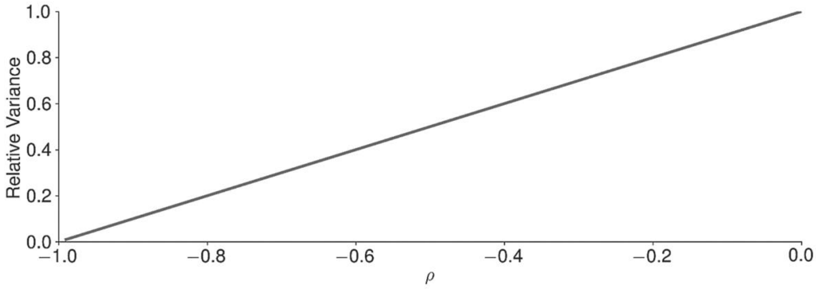
13.8 The most important limitation is that the bootstrap cannot generate data that has not occurred in the sample. This is sometimes called the "Black Swan" problem. A simulation can potentially generate shocks larger than historically observed values if the assumed distribution has this feature.

Solved Problems

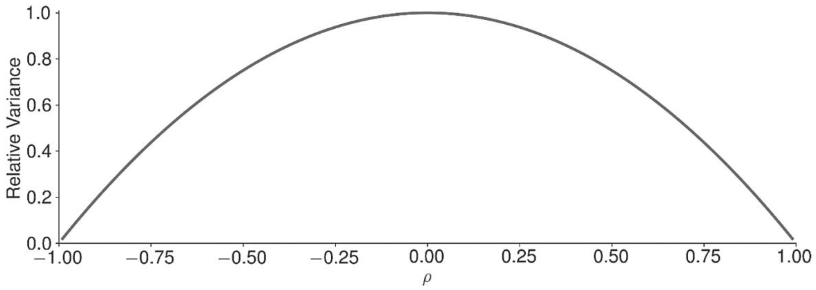
13.9 The standard deviation is USD 12.30, and a 95% confidence interval is $[\hat{\mu} - 1.96 \times \frac{12.30}{\sqrt{n}}, \hat{\mu} + 1.96 \times \frac{12.30}{\sqrt{n}}]$ and so the width is $2 \times 1.96 \times \frac{12.30}{\sqrt{n}}$. If we want this value to be 1% of USD 47.00, then $2 \times 1.96 \times \frac{12.30}{\sqrt{n}} = 0.47 \Rightarrow \sqrt{n} = \frac{2 \times 1.96 \times 12.30}{0.47} = 102.5$ (so 103 replications). Using 0.1%, we would need 1,025.8 (replace 0.47 with 0.047) and so 1,026 replications.

The following questions are intended to help candidates understand the material. They are *not* actual FRM exam questions.

13.10



13.11



13.12 Essentially, the question is to determine the cumulative distribution up to 1.2 on an N(1,5) distribution.

This can be done with the Excel function NORMDIST(1.2,1,sqrt(5),True) and returns the value 0.535.

13.13 **a.** The formula is simply implemented in Excel. For the first entry:

Max(−0.65 − 1,0) = 0, etc.

	N(0,1)	Option Payoff
1	−0.65	0
2	−0.12	0
3	0.92	0.42
4	1.28	0.78
5	0.63	0.13
6	−1.98	0
7	0.22	0
8	0.4	0
9	0.86	0.36
10	1.74	1.24

b. Recall that for a N(0,1) distribution:

$$F_X(u_1) = -F_X(1 - u_1)$$

Accordingly, the table is populated as follows:

	N(0,1)	Option Payoff	Antithetic Sample	Option Payoff
1	−0.65	0	0.65	0.15
2	−0.12	0	0.12	0
3	0.92	0.42	−0.92	0
4	1.28	0.78	−1.28	0
5	0.63	0.13	−0.63	0
6	−1.98	0	1.98	1.48
7	0.22	0	−0.22	0
8	0.4	0	−0.4	0
9	0.86	0.36	−0.86	0
10	1.74	1.24	−1.74	0

c. Using Excel, the correlation between the option payoffs is −27%. Therefore, the reduction in error is:

$$1 - \sqrt{1 + \rho} = 15\%$$

For the same number of simulations.

d. The Max in the payoff function is

$$\text{Max}(|x - 0.5|,0) = |x - 0.5|$$

The table is now as follows:

	N(0,1)	Option Payoff	Antithetic Sample	Option Payoff
1	−0.65	1.15	0.65	0.15
2	−0.12	0.62	0.12	0.38
3	0.92	0.42	−0.92	1.42
4	1.28	0.78	−1.28	1.78
5	0.63	0.13	−0.63	1.13
6	−1.98	2.48	1.98	1.48
7	0.22	0.28	−0.22	0.72
8	0.4	0.1	−0.4	0.9
9	0.86	0.36	−0.86	1.36
10	1.74	1.24	−1.74	2.24

The correlation between the two option value vectors is +25%, so the INCREASE in standard error is $\sqrt{1 + \rho} - 1 = 12\%$.

INDEX

A

Akaike information criterion (AIC), 179, 180
annualized volatility, 215, 216
antithetic variables, 231
AR models. *See* autoregressive (AR) models
attenuation bias, 107
autocorrelation, 164
autocorrelation function (ACF), 164, 166, 171
autocovariance, 163–166
autoregressive conditional heteroskedasticity (ARCH) process, 165
autoregressive (AR) models, 162
  AR(*p*) model, 167–168
  autocovariances, 166
  default premium, 169–170
  lag operator, 166–167
  real GDP growth rate, 170, 171
  Yule-Walker equations, 169
autoregressive moving average (ARMA) processes, 173–176
axioms of probability, 4

B

Bayesian information criterion (BIC), 179–180
Bayes' rule, 7
Bernoulli distribution, 28–29
Bernoulli random variable, 13
best linear unbiased estimator (BLUE), 70–71, 149–150
best unbiased estimator (BUE), 149
Beta distribution, 40
bias-variance tradeoff, 143–144
binomial distribution, 29–30
bivariate confidence intervals, 131, 132
Black-Scholes-Merton model, 214, 216
BLUE. *See* best linear unbiased estimator (BLUE)
Bonferroni (or Holm-Bonferroni) correction, 93
bootstrapping, 233
  limitations, 236
  stock market returns, 234–236
  vs. Monte Carlo simulation, 236
Box-Jenkins methodology, 180

C

CAPM, 111–113
CBB. *See* circular block bootstrap (CBB)
central limit theorem (CLT), 71, 73, 75, 76, 228
central moments, 16
characteristic equations, 184
χ^2 (chi-squared) distribution, 36–37
circular block bootstrap (CBB), 233
  sample generation using, 234
  stock market returns, 234–236
cokurtosis, 76–78
common univariate random variables
  continuous random variables, 31–42
  discrete random variables, 28–31
conditional correlation coefficient, 127
conditional distributions, 50–51
conditional expectation, 55–56
conditional independence, 6–7, 55–56
conditional probability, 4–5
confidence intervals, 204
  hypothesis testing, 90–91
  multivariate, 130–133
constant variance of shocks, 108

continuous random variables, 18–20, 56
approximating discrete random variables, 34
Beta distribution, 40
χ^2 distribution, 36–37
exponential distribution, 39–40
F distribution, 38–39
log-normal distribution, 34–37
mixture distributions, 40–43
normal distribution, 32–34
Student's *t* distribution, 37–38
uniform distribution, 41–43
controls, 125
control variates, 231–232
Cook's distance, 149, 150
correlation, 53–55, 74–75
vs. dependence, 218–221
matrix, 221
coskewness, 76–78
covariance, 54–55, 74–75
covariance stationarity, 163–164
cumulative distribution function (CDF), 12

D

data-generating process (DGP), 219, 226
default premium, 169–170
dependent white noise, 165
discrete random variables, 13–14, 48–51
Bernoulli, 28–29
binomial, 29–30
conditional distributions, 50–51
independence, 50
marginal distributions, 49–50
Poisson, 30–31
probability matrices, 48–49
distribution of financial returns, 216–218
dummy variables, 104, 193

E

Eiker-White estimator, 144
events, 2–5
event space, 2–5
exchange-traded funds (ETFs), 113
expectation operator, 14
expectations, 51–52
expected value of a random variable, 14
exponential distribution, 39–40
extraneous variable, 143

F

false discovery rate (FDR), 93
Fama, E. F., 124
Fama-French three-factor model, 124
familywise error rate (FWER), 93
F distribution, 38–39
Feasible Generalized Least Squares (FGLS), 147
feasible weighted least squares, 147
Federal Reserve Economic Data (FRED) database, 67
Fidelity's Contrafund, 115
financial asset, 215, 216
first-order AR process, 162
forecasting
non-stationary time series, 203–204
stationary time series, 180–183
French, K. R., 124
F-test statistics, 129–131, 133
fundamental principles, 2
fund managers, 114

G

Gauss, C. F., 149
Gaussian distribution, 32
Gauss-Markov Theorem, 149
general-to-specific (GtS) model, 143

H

hedging, 113–114
heteroskedasticity, 144–147
higher order moments, 67–70
higher-order, non-central moments, 41
high-performing fund managers, 114
histograms, 69
hypothesis testing, 110–111
about the mean, 86–88
confidence intervals, 90–91
critical value and decision rule, 86
definition, 84
effect of sample size on power, 89–90
effect of test size and population mean on power, 90
null and alternative hypotheses, 84–85
one-sided alternative hypotheses, 85
p-value of test, 91–93
testing the equality of two means, 93–94
test statistic, 85
type I error and test size, 86
type II error and test power, 88–89

I

iid bootstrap, 233
sample generation using, 234
stock market returns, 234–236
iid random variables, 107–108
implied volatility, 216

independence, 5–7
  conditional, 6–7
independent, identically distributed random variables, 57–58
independent variables, 54–55
inference testing, 110–111
information criterion (IC)
  Akaike, 179, 180
  Bayesian, 179–180
interquartile range (IQR), 20
invertibility, 180

J

Japanese yen/US dollar rate (JPY/USD), 201, 202, 216, 217
Jarque-Bera test statistic, 217
Jensen's inequality, 15, 204

K

Kendall's τ, 219–221
Kernel density plots, 69
kurtosis, 16–17, 67–69
k-variate linear regression model, 122. *See also* multiple linear regression

L

law of large numbers (LLN), 71, 73
linear correlation, 218
linearity, 103
linear process, 162–163
linear regression, 102
  CAPM, 111–113
  dummy variables, 104
  hedging, 113–114
  inference and hypothesis testing, 110–111
  linearity, 103
  OLS, 104–110
  parameters, 102–103
  performance evaluation, 114–115
  transformations, 103–104
linear transformations, 17–18
linear unbiased estimator (LUE), 70
Ljung-Box *Q*-statistic, 195, 196
log-normal distribution, 34–37
log-quadratic time trend model, 191
log returns, 214, 215
LUE estimator, 149

M

marginal and conditional distributions, 57
marginal distributions, 49–50
market variation, 124
Markov, Andrey, 149
maximum likelihood estimators (MLEs), 150
mean, 16–17
  estimation of, 64–65
  of random variable, 228–229
  sample behavior of, 71–73
  scaling of, 66
  and standard deviation, 66–67
  of two variables, 75–76
  and variance using data, 67
median, 73–74
mixture distributions, 40–43
MLEs. *See* maximum likelihood estimators (MLEs)
model selection, 176–179
moments, 52–55
  multivariate (*See* multivariate moments)
  random variable
    central moments, 16
    definition, 14
    kurtosis, 16–17
    linear transformations, 17–18
    mean, 16–17
    non-central moments, 15, 16
    skewness, 16–17
    standard deviation, 16
    variance, 16–17
Monte Carlo simulation
  antithetic variables, 231
  approximating moments, 226, 227
  control variates, 231–232
  generating random values/variables, 226, 227
  improving accuracy of, 231
  limitation, 233
  mean of random variable, 228–229
  price of European call option, 229–231
  vs. bootstrapping, 236
moving average (MA) models, 171–173
multicollinearity
  outliers, 147–149
  residual plots, 148
multi-factor risk models, 124–127
multiple explanatory variables, 122–124
multiple linear regression, 103
  F-test statistics, 129–133
  with indistinct variables, 122
  model fit measurement, 127–129
  multi-factor risk models, 124–127
  with multiple explanatory variables, 122–124
  multivariate confidence intervals, 130–133
  R^2 method, 128–129
  stepwise procedure, 123
  testing parameters, 129–133
  t-test statistics, 129

multivariate confidence intervals, 130–133
multivariate moments, 74
coskewness and cokurtosis, 76–78
covariance and correlation, 74–75
sample mean of two variables, 75–76
multivariate random variables
conditional expectation, 55–56
continuous random variables, 56–57
discrete random variables, 48–51
expectations and moments, 51–55
independent, identically distributed random variables, 57–58

N

non-central moments, 15, 16
non-stationary time series
cyclical component, 194
default premium, 202, 203
forecasting, 203–204
log of gasoline consumption, in United States, 194–195
random walks and unit roots, 196–201
seasonality, 192–194
spurious regression, 201, 202
time trends, 190–192
normal distribution, 32–34
null and alternative hypotheses, 84–85

O

OLS estimators. *See* ordinary least squares (OLS) estimators
omitted variables, 107, 142–143
one-sided alternative hypotheses, 85
ordinary least squares (OLS), 104–107
constant variance of shocks, 108
iid random variables, 107–108
implications of assumptions, 109–110
no outliers, 108–109
parameter estimators properties, 107–110
shocks are mean zero conditional, 106–107
standard error estimation, 110
variance of *X*, 108
ordinary least squares (OLS) estimators, 122, 123, 128
homoscedasticity, 144, 147
strengths of, 149–151
weighted least squares, 147

P

PACF. *See* partial autocorrelation function (PACF)
parsimony, 180
partial autocorrelation function (PACF), 164, 166–168, 171, 172, 176–180
Pearson's correlation, 218
performance evaluation, linear regression, 114–115
PIMCO High Yield Fund, 115
PIMCO Total Return Fund, 115
Poisson random variables, 30–31
polynomial trends, 190–192
portfolio diversification, 53–54
power laws, 217–218
probability
Bayes' rule, 7
events, 2–5
event space, 2–5
fundamental principles, 2, 4
independence, 5–7
matrices, 48–49
in rectangular region, 56
sample space, 2–5
probability mass function (PMF), 12, 13
pseudo-random number generators (PRNGs), 226
pseudo-random numbers, 226
p-Value-at-Risk (*p*-VaR), 234
p-value of test statistic, 91–93

R

random variables
continuous, 18–20
cumulative distribution function, 12
definition of, 12–13
discrete, 13–14
expectations, 14
modes, 20–22
moments, 14–18
probability mass function, 12
quantiles, 20–22
random walks, 196–201
rank correlation, 218–221
regression
analysis, 102
with multiple explanatory variables (*See* multiple linear regression)
regression diagnostics
bias-variance tradeoff, 143–144
extraneous variable, 143
heteroskedasticity, 144–147
multicollinearity, 147–149
omitted variable, 142–143
returns, measuring, 214
R^2 method, 128–129

S

sample autocorrelation, 176–179
sample moments, 64
best linear unbiased estimator, 70–71
higher order moments, 67–70
mean and standard deviation, 66–67
mean and variance using data, 67
mean estimation, 64–65

median, 73–74
multivariate moments, 74–78
sample behavior of mean, 71–73
variance and standard deviation estimation, 65–66
sample selection bias, 106
sample space, 2–5
seasonal differencing, 200, 201
seasonality
non-stationary time series, 192–194
stationary time series, 183–184
simple returns, 214, 215
simulation, 226
conducting experiments, 226, 228
Monte Carlo (*See* Monte Carlo simulation)
simultaneity bias, 107
skewness, 16–17, 67–70
Spearman correlation, 219–221
spurious regression, 201, 202
standard deviation, 16
mean and, 66–67
scaling of, 66
and variance estimation, 65–66
vs. standard errors, 66
standard errors
estimation, OLS, 110
standardized Student's *t* distribution, 37
stationary time series
ARMA models, 173–176
autoregressive models, 165–171
covariance stationarity, 163–164
forecasting, 180–183
model selection, 176–179
moving average models, 171–173
sample autocorrelation, 176–179
seasonality, 183–184
stochastic processes, 162–163
white noise, 164–165
stochastic processes, 162–163
Student's *t* distribution, 37–38
survivorship bias, 106
systemically important financial institutions (SIFIs), 5

T

test statistics, 85, 91–93, 110, 111, 130, 131, 133, 145–147, 176, 178, 193, 198, 199, 204, 217
time-series analysis, 162
time trends, 190–192
total sum of squares (TSS), 127, 128
transformations, linear regression, 103–104
t-statistics, 124, 125, 133
t-test statistics, 129
type I error and test size, 86
type II error and test power, 88–89

U

uniform distribution, 41–43
unit roots
problems with, 197
testing for, 197–200
univariate random variables, 12

V

validation set, 144
variance, 16–17
and standard deviation estimation, 65–66
of *X*, 108
VIX Index, 216
volatility and risk measures, 215–216

W

weighted least squares (WLS), 147
White's estimator, 144–146
Wold's theorem, 165

X

χ^2 (chi-squared) distribution, 36–37

Y

Yule-Walker (YW) equations, 169